FUNDAMENTALS *of* STATISTICS

BY

TRUMAN LEE KELLEY

HARVARD UNIVERSITY

HARVARD UNIVERSITY PRESS

CAMBRIDGE, MASSACHUSETTS

1947

Printed in the United States of America

To

EXPERIMENTALISTS

Though the mists of life
obscure the distant scene
they are undismayed for,
in markings near about,
they discern the contour
of the land and the portal
to the future.

PREFACE

An alternative title for this book which would emphasize the point of view is *Statistics, Its Philosophy and Method*. This work is more than a revision of the writer's earlier *Statistical Method* (1923). That work covered a field which in the last twenty years has grown to such magnitude that nothing short of an encyclopedia could cover it today. The endeavor herein has been to place a great emphasis upon the logic and principles underlying the statistical study of phenomena, to provide, in the early chapters, such basic issues as will integrate thoughtful and investigative moods with statistical processes, and, in the later chapters, to give such treatment of modern processes as is required in handling many experimental situations and as will open to the reader the wealth of thought in current statistical literature.

The early chapters constitute an elementary text, and call for no mathematical background other than that of arithmetic and elementary algebra. They seek, first, to show the place of statistics in the social process of solving problems and, second, to provide the most basic tools. The approach to statistics of most beginning students is that it holds some useful

tricks which, with a certain drudgery, can be learned. The content of statistics is not a bag of tricks, but rational thought processes which serve in problem situations. That there is a nicety in these processes gives zest to their discovery and pleasure in their use. This viewpoint can scarcely be overemphasized, and it is hoped that teachers of this text will further emphasize and elaborate upon the utility of statistical concepts in the everyday problems of the physical and biological scientist, the schoolman, the doctor, the farmer, the economist, and the businessman.

The parallel work of the author, *Statistical Tables*, is designed specifically for laboratory statisticians and is not a requisite to students of this text in view of the abridged set of tables given herein.

In addition to the numerous acknowledgments cited throughout this text, the writer is deeply indebted to Eric F. Gardner for statistical research in connection with sequential analysis and otherwise and he is also greatly indebted to him and to Katherine Ytredal for expert assistance in the difficult task of preparation of tables and of the manuscript for the printer.

Truman Lee Kelley

Cambridge, Massachusetts

CONTENTS

I. THE DIGNITY OF DATA AND THE BACKGROUND OF STATISTICS

II. STATISTICAL SERIES

III. STATISTICAL TABLES

IV. GRAPHIC METHODS

V. THE STABLE FEATURES OF PHENOMENA

VI. MEASURES OF VARIABILITY

VII. MEASURES OF CENTRAL TENDENCY

VIII. THE NORMAL DISTRIBUTION

IX. THE STATISTICS OF ATTRIBUTES

X. ESTIMATION, REGRESSION, AND CORRELATION

XI. FURTHER CORRELATION ISSUES

XII. THE GENERAL MULTIPLE LINEAR REGRESSION PROBLEM

XIII. SUNDRY STATISTICAL ISSUES AND PROCEDURES

XIV. MATHEMATICS, THE MENTOR OF STATISTICAL INGENUITY

XV. STATISTICAL TABLES

APPENDIX A. MATHEMATICAL BACKGROUND TEST

APPENDIX B. REFERENCE LISTS

FUNDAMENTALS OF STATISTICS

CHAPTER 1

THE DIGNITY OF DATA AND THE BACKGROUND OF STATISTICS

SECTION 1
STATISTICS, PSYCHOLOGY, AND LOGIC

An isolated fact is an unthinkable phenomenon. Adults experience them and think of them after the occurrence, but not before. Furthur experiences of the same sort (the impossibility of such in an absolute sense is admitted, for a second experience cannot be exactly the same as a first) are no longer isolated. Infants presumably experience them in great number, making each new day richer than the past in ability to interpret. These new infringements upon consciousness are a part of life which lies outside the field of statistics. Note the appropriateness of the plural, and that "statistic" is capable of definition only after the plural has been defined. As a second experience of something, say a cat caterwauling, is different from the first, so the third is different from the second, etc., *ad infinitum.* There is no identity in successive experiences. In some sense each is isolated from all others, but in another and a psychological sense it is not, and it is this latter fact that provides the basis for statistics.

Let us attempt to trace the delimitation of attention of a person in an investigative mood. We, of course, cannot catch him at the beginning of any activity. At the time we first see him he is already possessed of many ideas, experiences, remembrances, interests, and attitudes, dating from the more or less remote past, and he is in a particular environment providing particular problems and data. The nature of the investigation that he undertakes is limited by these things. If on a country road, and if in his background there is the training, the interest, and the specific information of a botanist, he may narrow his attention to the botanical features around, whereas if a geologist he may turn his attention to rocks and soils. This narrowing of the field of interest is necessary to contemplative and creative thinking. In statistical language it is the first step in the making of a statistical series. The items of such a series do not encompass all that exists, or even all that is knowable, about the things observed, but only so much about each as is apparent when the attention is limited to a specified narrow field.

*A statistical series is a succession of observed values referring to a single character of a number of objects, things, or cases, which have been selected according to some single principle. In the situation wherein these cases are randomly drawn members of a much larger, generally infinite, parent population, the series is also a sample.**

The establishment of the single principle of selection is the first step in methodical in-

* For an excellent discussion of sampling see A. L. Bowley, ELEMENTS OF STATISTICS, fifth ed., 1926, and also James G. Smith and Acheson J. Duncan, SAMPLING STATISTICS AND APPLICATIONS which is vol. 2 of FUNDAMENTALS OF THE THEORY OF STATISTICS, 1945.

vestigation. One cannot and should not try to investigate everything at once. To attempt it is scatterbrained and unproductive. The original limits of interest of the botanist on the country road might be to living things and thence to vegetation. A further limitation might be to trees. A still further one to maple trees, and thence to one characteristic of them, say leaves, and thence to, say, the veining of leaves. If, after following this greater and greater delimitation to a certain point, the investigator stops and then observes that there remains a variable phenomenon, say the number of veins, of a number of things, leaves, all falling within the limits of selection as he has defined them, he then has a statistical series, which in this illustration is also a sample, for the leaves noted may be thought of as a finite random selection from an infinite population of maple leaves.

Had he stopped earlier in the process of delimitation, say at the point when his interest was merely limited to living things, would he then have had a statistical series? The items might be the height of a pine tree, the bark of a dog, the speed of a pony, etc. Though these be under a single principle,—living things as encountered along the highway,—it is not enough, for there is also needed a single principle of observation. Had he observed the height of a pine tree, the height of a dog, of a pony, etc., we would call it a statistical series and also a sample, intractable though it may be in revealing any interesting or unifying fact of life.

If we attempt to define statistical series and samples more narrowly than stated we run into logical difficulties. If we require that the elements of a sample shall be in truth observations of a single sort, made upon things of a single sort selected in some single manner, we have narrowed it to the point of impossibility

of exact fulfillment. What are phenomena of a single sort and what is selection of a single sort?

Let us consider a sample coming about as closely as we can conceive to meeting these conditions,—the lengths of clam shells collected upon a certain beach. Are the recorded lengths the lengths of a single phenomenon? Is the selection of the successive shells made in a single manner? Are the shells clam shells? If the characteristics of two species, say clams and snails, are so different that there is no overlapping in regard to some character, we can certainly avoid getting snail shells when collecting clam shells, but many a problem of practical statistics is not of this sort. It may well be that some shells other than those originally thought of or defined as clam shells may be included in the sample. In the last analysis the safeguard against this lies in someone's judgment that shell *A*, all things considered, is of the same species as shell *B*. We may call this crucial act *the judgment of sameness*.

Next consider length,—what is the length of a clam shell? Here is one which is chipped at the end and we discard it, but here is one with reference to which it is hard to say whether it has been chipped or not, and we are again forced to depend upon a judgment of sameness.

Now consider the selection of shells under similar conditions. Shell *A* is collected at a certain time and space, and this identical time and space does not provide us with another shell. Shell *A* is collected at high, and shell *B* at low tide. Are these collected in some single manner? Clearly no. All we can ever assert is that the items have been collected under conditions judged the same so far as factors that we judge material are concerned. Stated otherwise, there has been a *judgment of irrelevance* of those conditions

of sampling that have been observed to vary from item to item. Under optimum conditions we can assert that the collector's threshold of difference recognition, as it pertains to the conditions of sampling, is not exceeded by the several experiences yielding the observations of his sample, but we may not claim more than this. The logical issues connected with sampling, probability, and truth have been thoroughly discussed by von Mises (1939).

Mathematical statistics can sidestep all these issues by simply assuming that a true sample is at hand. This is never the case. To realize that applied statistics does not rest upon pure logic may be a disappointment to the reader, but it is because it does not that it concerns itself with phenomena, not noumena, and that it is adaptable to all the problems of life in their partly repetitive aspects, which are their chief aspects. All the fine-spun deductions of mathematical statistics are by some amount wide of the mark when applied to real phenomena. Logically this must be so. They are astray by an amount which judgment alone bears witness to. How wide and what of it are questions that are very difficult of quantitative answer. As these difficulties are analyzed they regularly run back to the question of the soundness of some judgment of sameness or of relevance.

Consider the last. The geologist finds a certain fossil at 10 a.m. and a second at 3 p.m. of a certain day. It may not occur to him that the time of finding is pertinent to any intrinsic characteristic of the fossils and he neglects mentioning the time. Whether overt or not, this is a judgment that the time is immaterial. The non-overt judgments of the novice frequently lead to insidious and serious error. If

the first specimen is found in a gravel bed and the second in a clay bank, he undoubtedly would consider these space conditions as material and he would not report the items as of a single sample. The effective judgment may thus be either a positive or a negative act. Sampling errors due to this latter are very insidious and are committed by the man of narrow vision who blandly steps in where wiser men refrain to tread.

The hazard of the positive act of judgment, which the collector of a sample makes, is of a very different and lesser sort. The judgment of relevance, if made consciously and defended, may be one of those judgments which mankind can make with considerable competence. Consider the following demands upon judgment, assuming no instrumental aids in making the judgments: (a) is *A* twice as tall as *B*? (b) is *A* 1.5 times as tall as *B*? (c) is *A* 1.1 times as tall as *B*? (d) is *A* just as tall as *B*? The accuracy of the respective judgments is a matter of psychology, which would probably inform us that the accuracy of the sameness judgment is far greater than that of any of the other judgments. Certainly in the matter of pitch discrimination a very small difference, say that between 256 and 256.1 vibrations a second, can be distinguished when the two sounds are equally loud and heard under equally favorable conditions and differ in time from each other by a very short interval. Here the query is, "Are the two the same?" and the answer can be given with a precision far exceeding the judgment that *A* is some multiple of *B*. Though we lack a general proof we may well believe that the sameness judgment is the most precise judgment within the power of the human mind to make.

In view of this we need not be unduly dis-

turbed by the fact that the pertinence of the beautiful techniques of mathematical statistics to an actual problem are contingent upon some earlier act or acts of judgment. There is a real and important difference between mathematical statistics and the applied statistics with which the scientist is concerned in his attempt to discover or verify principles and laws underlying observable phenomena.

There is a quadruple alliance inherent in sound statistical research: phenomena, i.e., data; logic, i.e., mathematics; human psychology in its power to judge sameness; and human psychology in its power to appraise relevance. If this total activity is to be at its best, the data must be germinal, the mathematical elaboration sound and penetrating, the judgment of sameness made with reference to things that the mind has primary faculties for sensing, and the judgment of relevance made in the light of an imagination so rich that few possibilities and connections are overlooked.

We may not assert with reference to any data collected to throw light upon an unsolved issue that it merely awaits the expert to bring to fruit the important germ within. Though good judgment determined the data selected, it may still be barren. Also, prior to study, we cannot assert that any new data consequent to causes not known a priori, are void of virtue, but one large class does seem peculiarly sterile. *In dealing with biological and social phenomena samples not composed of cases competitive with reference to the issues studied are, characteristically, barren of important relationships.*

SECTION 2. HISTORICAL ANTECEDENTS OF MODERN STATISTICS

Modern statistics is the outgrowth of a number of lines of social and scientific development. Because of this mixed ancestry, and because the convergence of these lines into a single discipline, statistics, is none too obvious, there are strong lines of cleavage in present processes, and even in avowal as to the function of statistical method. Historically we find trade and vital statistics developing with the growth of business, particularly of international trade, and with national concern for welfare of citizens, their availability for the army, and their racial and vocational characteristics. The comprehensive work *The History of Statistics* (1918), edited by John Koren, is an excellent picture of this growth. The topics treated of in this work indicate the extent and detail reached through this line of development: Here are a few only of the topics falling under the first four letters of the alphabet under the general heading "United States."

Accidents
Agriculture census
Banks, statistics of
Birth statistics
Census Office
Children's Bureau
Church statistics
Civil War Veterans
Coal trade
Commerce Commission,
  Interstate
Commerce, Department of
Commerce foreign
Commerce internal
Corporations
Cotton
Crime
Crop Estimates
Death statistics
Defective classes
Delinquency
Divorce statistics
Duties

Social phenomena have teemed with classifiable facts, and the student, whether tradesman or legislator, has realized that the careful collection of these facts, their assorting into independent or related categories, for successive intervals of time, was informative of the course of events, enabling an understanding and control of them not otherwise possible. This line of statistical development may be characterized as a response to the welling of social phenomena. With this as a starting point the technique of the accountant in recording transactions in money or things of monetary value is taken over and adapted to the recording of amounts in categories other than monetary. The good bookkeeper becomes the good statistician, and that is so.

But this is only part of the story of modern statistics. Let us turn to *Studies in the History of Statistical Method*, by Helen M. Walker, (1929). The prospectus of this book states that it "presents the modern use of statistics against a background of the work of De Moivre, Bernoulli, Gauss, Laplace, Quetelet, Galton, Ebbinghaus, Fechner, and many others." Here we find different names, different topics, and a radically different emphasis, as witness the following entries occurring under the first four letters of the alphabet in the table of contents:

Airy
Anthropologists, statistical work of
Association
Astronomers, contribution to the theory of probability
Attributes
Average
Bernoulli, Jacques
Bernoulli, N.
Bessel
Binomial expansion
Boas
Bowley
Bravais
Charlier
Chi-square
Contingency
Correlation
Correlation, partial
Correlation, multiple
Curvilinear regression
Differences
Distribution

A most elementary acquaintance with the names and topics here listed shows a vast difference in approach from that previously mentioned. Modern statistics is a very real fusing of varied antecedents which can be analyzed into much greater detail than into just the two particular lines mentioned, which we may call the social and the scientific backgrounds. In the scientific background we find developments in each of the sciences progressing more or less independently of each other. There is a statistics of economics, of life insurance, of biology, of psychology, of engineering, of physics and chemistry, of astronomy, and of mathematics. The developments in each of these fields are continually running into related fields, but each has its taint, such that a trained statistician unfamiliar with a given foreign language could almost certainly pick up a statistical work written in that language and allocate it to its appropriate field merely by observing the formulas upon the pages.

SECTION 3. OCCASIONS FOR RESORT TO STATISTICS

Differences in logical antecedents: A very fundamental difference depends upon an underlying difference in the logical issue that is set. We may say that there are two occasions for resort to statistical procedure, the one dominated by a desire to prove a hypothesis, and the other by a desire to invent one. This has led to distinct schools of statisticians, both lying within the general field of scientific endeavor.

The first school is that represented by mathematicians who start with certain elementary principles and deduce therefrom facts of distribution, frequency, and relationship. In so far as observed situations parallel these conclusions the same elementary principles are supported as applying to the data in hand. One weakness of

this approach lies in the fact that a number of causes — different sets of elementary principles — may result in substantially the same net result. A still greater weakness is that it is essentially a deductive procedure and relatively sterile in suggesting new causes — in inspiring creative inferences. It is fundamentally a method of proof and not one of invention; and just because it is a method of proof, it has a permanent place in statistical method. It must, however, if in the service of the social and biological sciences, be but a handmaid to the creative genius of mathematical synthesis and induction.

The second school is best represented by those biometricians and economists who start with observed data and endeavor so to group them and treat them that the constant features of the data are made apparent. This is a process of statistical analysis. It may at times be expected to be an involved process, for social phenomena are complex. Data are frequently warped to fit statistical convenience, but if statistics is to realize its high destiny, procedure must be flexible, for only when the method is mobile can it fit immobile data. The accurate measurement of those features of phenomena which are exceptional, but not chance, is the unique province of statistical analysis.

The present treatment follows the second school, for the starting point of every problem is data, actual data and not hypothetical. When data are found to approach a hypothetical form, as for example at times normality of distribution, then the beautiful relationships of the theoretical or ideal form are availed of to the utmost, but it is intended that the treatment will never lose the characteristic of its origin, and that an observed agreement with objective

phenomena will always be considered proof of soundness of the technique pursued.

The method of approach in this text is inductive, starting with data and deriving constants, and will not give the noumenal satisfaction that comes from tossing coins, throwing dice, and sorting cards, thus obtaining distributions which approach an ideal standard.

Psychological warrant for statistics: It is probably true that in the social experiment of living difficult procedures are a sort of last resort. If the issues of life can be solved by logic, why use any other method? If the issues can be solved by the nearly exact procedures of chemical or physical analysis, why go to the bother of dealing with many samples? If the biological laws of inheritance are given by knowable and unalterable laws of cause and effect. why investigate general tendencies of mating and of family relationship? In each instance there is no sufficient reason. The facts, however, seem to show that life does not reveal itself with such certainty. Even in chemistry and physics, only approximate answers to problems are given in exact and concise terms,—the inexactness in observed fulfillment being due not only to "impurity" of the data at hand, but also to the fact that logical and mathematical formulations are characteristically, probably universally, but approximations to reality.

Where the approximations are close and where the data are, or can by selection or experimental refinement be made, relatively pure. the adequacy of the "principle," the "law," and the "formula" is greatest and statistical techniques are in least demand. Technology well represents this field as does that upon which it is built, the "established" facts of science. However, in the process of establishing such facts a very differ-

ent attitude was present. The physicist observes seemingly irregular changes in X as Y changes. He repeats his experiment, controlling more and more of the conditions, and repeats again and again, and, if successful, reaches a law at the end of his work. He has been using statistics, though the engineer who utilizes his law does not. Statistics are inherently connected with the formulative stages of knowledge. This is true with regard to statistics as accumulations of repeated observations, and as techniques for discovering the characteristic features of such accumulations. Statistics and experimental research are wedded. In the case of the physicist introducing one after another greater degree of control over his data, it is especially to be noted that the stimulus to do so has been consequent to antecedent statistics, i.e., variability in several or many observations in which variability was not expected. In cutting down this variability through the greater control of conditions, he may say that he is avoiding statistical techniques, but the fact is that statistical considerations have been the very cause of these later steps. In one sense we may say that the main purpose of statistics has been to eliminate itself. This is a sound view. Statistics is a research instrument and drops out of the picture to the degree that the problem becomes solved. However, as long as undetermined variability in outcome is present, the statistical aspects of the situation remain. In this sense most of the important problems of social life are never solved, but are only in the process of being solved.

This can be adequately expressed in terms of the total and the part variances entering into a statistical situation. We will illustrate the meaning of variance with the 20 hypothetical temperatures at a single weather station, as listed in Table I A.

TABLE I A

TEMPERATURES			Deviations from Means			Squares of Deviations			Products
Actual	Predicted	Errors							
X	$\bar{X}$	e	x	$\bar{x}$	e	x^2	$\bar{x}^2$	e^2	$\bar{x}e$
74	73	1	4	3	1	16	9	1	3
74	74	0	4	4	0	16	16	0	0
66	66	0	-4	-4	0	16	16	0	0
76	73	3	6	3	3	36	9	9	9
70	73	-3	0	3	-3	0	9	9	-9
72	73	-1	2	3	-1	4	9	1	-3
66	67	-1	-4	-3	-1	16	9	1	3
74	72	2	4	2	2	16	4	4	4
70	72	-2	0	2	-2	0	4	4	-4
70	67	3	0	-3	3	0	9	9	-9
72	72	0	2	2	0	4	4	0	0
70	68	2	0	-2	2	0	4	4	-4
64	67	-3	-6	-3	-3	36	9	9	9
68	67	1	-2	-3	1	4	9	1	-3
68	70	-2	-2	0	-2	4	0	4	0
66	68	-2	-4	-2	-2	16	4	4	4
72	70	2	2	0	2	4	0	4	0
68	68	0	-2	-2	0	4	4	0	0
72	70	2	2	0	2	4	0	4	0
68	70	-2	-2	0	-2	4	0	4	0
Sums 1400	1400	0	0	0	0	200	128	72	0
Means 70	70	0	0	0	0	10.0	6.4	3.6	0

The predicted values are those made by the forecaster 24 hours earlier and based upon all the meteorological information available to him at that time. Thus $\bar{X}$ is that which is known about the situation and e is that which is unknown. For each item the actual temperature, X, equals the predicted temperature, $\bar{X}$, plus the error of prediction, e, i.e., $X = \bar{X} + e$. The same holds when dealing with deviations from means, for $x = X - M$, and $\bar{x} = \bar{X} - M$, and $e = e - 0$, so that $x = \bar{x} + e$.

The quantity e, being the error in the prediction, is in the illustrative data of Table I A, as it universally must be, independent of the prediction $\bar{x}$. Otherwise expressed, as later proven in connection with the derivation of a correlation coefficient, the sum of the $\bar{x}e$ products is equal to zero. In this case the variance of the $\bar{x}$'s (their mean square) plus the variance of the e's exactly equals the variance of the x's. We note that 10.0 = 6.4 + 3.6 and in general $V_x = V_{\bar{x}} + V_e$. This important property of divisibility into independent additive parts of this particular measure (the variance) of variability is not a property of any other measure of variability and is the basic reason why this measure facilitates thinking as does no other measure of variability.

Let us choose such units that the variance in temperatures when no defining facts are given is 1 (not 10 as in the data of Table I A). If the defining facts give such information that it is now known that the range of probable temperatures has a variance ½, there is still much variability in the temperature to be expected 24 hours hence. The forecasting formula is inadequate to give precise knowledge, so statistical issues of distribution of probable temperatures remain. They will continue to remain until forecasting is more

accurate and the person, whether meteorologist or layman, interested in this residual 50 per cent of the variance will continue indefinitely to be concerned with a statistical problem. Table I B may help to picture the field wherein statistics operates.

TABLE I B

SITUATIONS CLASSIFIED ON BASIS OF USE OF STATISTICS AND STATISTICAL METHODS

Initial Variance	Knowledge of factors determining it	Residual variance after using such knowledge	Resulting in Situations wherein	
			Statistics are commonly not used by	Statistics are used by
1.000	Highly precise	.001	A. Chemists, engineers, and classes B, C, D, and E.	F. Research scientists in certain fields of astronomy, physics, bureau of standards, coast and geodetic survey, etc.
1.00	Excellent	.05	B. Technologists, "practical" professional men, and classes C, D, and E.	Class F and G. Research technologists, scientists in such fields as the physical sciences.

TABLE I B

(CONTINUED)

Initial Variance	Knowledge of factors determining it	Residual Variance after using such Knowledge	Resulting in Situations wherein	
			Statistics are commonly not used by	Statistics are used by
1.00	Very Useful	.20	C. Rule of thumb workers executives, men who "do things," school men, and Classes D and E.	Classes F, G, and H. Research workers in medicine, education, psychology, the biological sciences, and occasionally in the social sciences.
1.00	Useful	.50	D. Promoters, uncritical people, and Class E.	Classes F, G, H, and I. Research workers in the social sciences, the more careful executives, salesmen, guidance counselors
1.00	Meaningful, but poor	.80	E. Semi-competent "students," charlatans, and uncritical enthusiasts, gamblers who characteristically "foot the bill."	Classes F, G, H, I, & J. Fairly intelligent people in all walks of life. Words indicative of such use are "probability," "tendency," "likelihood," "usually," "sometimes," "generally," etc.

The primary field of statistics is represented by the middle portion of the table. When knowledge is nearly complete, and .999 of the variance can be accounted for, but few issues hinge upon the small fraction of the variance which is unknown. When knowledge is very inadequate, the entire field is avoided so far as possible by serious people, and especially by scientific workers, and is left in the hands of uncritical people. The occasion for statistical thinking is so common that the question is not whether one shall use it, but how fine a tool it is in his hands. This book aims merely to make explicit certain logical implications of phenomena, and in no sense to provide a tool that has not first been demanded by the requirements of clear thinking, as applied to quantitative and qualitative facts of life.

It should be obvious from Table IB that statistical methods should serve well many of the common needs of ordinary folk. However, their use and value in this connection is limited by understanding and not by the intrinsic fitness of the tool. In writing for a lay audience certain definite restrictions as to concepts and techniques employed are imposed. These may be very serious, making it well-nigh impossible to illustrate crucial points. Such terms as *variance* and *residual*, used in preceding paragraphs, are taboo. Substitution for the word *residual* can be made, but no substitution for *variance* is possible without first taking the time to teach quite a bit of statistics. As a consequence a large and important class of problems wherein a total variance is capable of analysis into its component parts is beyond the scope of the understanding of a popular audience. Undoubtedly attempts at popular presentation of this concept will be made, but we may expect them to be as unsatisfactory as the attempt to explain corre-

lation by means of such common concepts as percentages of a total, proportion of agreement, and other elementary mathematical concepts, all of which give the impression of conveying information while actually leading astray. What does it mean to say that a correlation is 66 per cent of perfect, or such that there is 66 per cent agreement? Presentation of such concepts to lay audiences is fraught with misunderstanding.

There remain many statistical concepts and techniques which may be employed in this case: There are available certain graphic methods, time charts, circle charts, growth curves, frequency polygons, means, ratios, percentages, limits, ranges, categories, frequencies in classes, and still other concepts. Some of these have been sadly overworked, as for example the "ratio" in connection with I.Q.'s (intelligence quotients), E.Q.'s (educational quotients), A.Q.'s (accomplishment quotients), etc., where concepts of growth, dispersion, and correlation are statistically more neat and informative. It should not be assumed that with literary skill the most recondite concepts can be presented to any audience of normal intelligence. This is scarcely so. Have the valient attempts to present "relativity" in the Sunday supplements resulted in a nation informed upon this subject? The next three chapters are concerned with such elementary statistical techniques that they are serviceable in connection with popular presentation, but it must not be assumed that they reasonably circumscribe the field of statistical effort of the practical school man, psychologist, or sociologist. They merely serve as an introduction to quantitative thinking.

Analysis of phenomena: Papers presented to scientific bodies may be concerned with the analysis of phenomena, but presentations to popular audiences are primarily propaganda,—legiti-

mate enough if the statistical picture given is accurate, but clearly serving a noninvestigative and nonanalytical purpose. The subtler phases of statistics deal with analysis, not with description. It is the good fortune of propagandists that this is so. It is very desirable that such a one should not overstep the limitations of his technique. To present some simple time chart and, by noting its ups and downs, proceed to analyze it and explain underlying causes is chicanery or wishful thinking. The analysis of a time chart calls for information far beyond that given in the chart itself. It is very easy for an investigator without dishonest intentions to (1) have a theory, (2) draw a chart, (3) see tendencies in it which are explicable in harmony with his theory, (4) think that the chart supports his theory, (5) present it to an audience, and (6) deduce in the presence of the audience the theory as consequent to the phenomena of the chart. Procedures like this discredit statistics and have, in poorly informed quarters, led to the charge that one can prove anything by statistics.

There is but a single story inherent in a given body of data and this can be approximated to by the delicate tool of statistical analysis. Its uniqueness is its outstanding feature. To believe, or try to get, several stories out of it, or merely proof for some preconceived story, is denying its genius. A Turner might paint the Smoky Mountains in brilliant red and present a charming picture, but what a sad story this would be to one who knows them, and what a false guide to the scout seeking landmarks in territory new to him.

The first function of statistics is to be purely descriptive, and its second function is to enable analysis in harmony with hypothesis, and its third function to suggest by the force

of its virgin data analyses not earlier thought of. In carrying out the first function we shall deal in forthcoming chapters with graphic portrayal, phenomena of distribution of a single variable, and phenomena of relationships between two or more variables. All of this is elementary statistics.

More advanced statistics concerns itself with such a study of given phenomena that invariant features of it are discovered,—they being there, of course, all the time, merely waiting an appropriate throwing aside of overlaid rubble to bring them to light. If a brilliant hypothesis has correctly conceived of some feature as being invariant, the testing out of the hypothesis is generally not difficult though transcending the simplicity of mere description. If no such hypothesis is present, there still is a possibility, by viewing the data now from this aspect, now from that, of discovering its stable features. If charged with painting a mountain, an artist might first ride around it on horseback to discover "the" vantage point above all others. So the statistician with new material may search diligently and by devious methods to find the unique story in his data. All the interest of the detective, all the joys of the artist, may be combined in his search.

Though statistical analysis has much in common with pure mathematics, there is one important difference. The mathematician assumes his initial conditions and then searches out their necessary consequences, building a more and more beautiful and intricate mathematical structure. The statistician starts not with an assumption, but with given data with their conditions denying arbitrary assumptions, and seeks to build a thought structure in harmony with it, and to indicate consequences that are in harmony with it. If the statistician could but first choose the nature of

his data and then ramify at heart's content and announce consequential relationships, more and more remote consequences could be pointed out. But since given data are inflexible in nature, then as characteristics are observed and deductions made, the more remote the deduction, the more uncertain its verity. Remote deductions are like the end of a whiplash. It may oscillate violently, though the whip handle move but slightly. The argument involving a long chain of consequential relationships which is so powerful in mathematical analysis can seldom be employed in statistical analysis. The slight initial error always to be expected in finite data may become augmented into a grievous falsehood by the time the end of the long argument has been reached. *For sound statistical analyses the magnitude of errors in the initial starting point must be more or less accurately apprehended and their significance carried along through every step so that the final outcome of an analysis has attached to it a definite concept of probable error.* The process of attaching to each step of a deductive argument a correlative step of probable error may be long and difficult, but it can scarcely be abridged.

If now we turn to mathematics as a tool in statistics, we shall find it essential and of universal service.

SECTION 4. THE MATHEMATICAL BACKGROUND OF STUDENTS OF STATISTICS

The common observation of teachers of statistics that their students are ill-prepared in arithmetic as well as algebra arouses no great surprise in colleagues in other departments, because every teacher is prone to blame the shortcomings of his instruction upon the earlier training of his pupils. Nevertheless, the teacher of elementary statistics has surprising

ground for his charge. The writer's experience with elementary students has mainly been with those in education and psychology, but it has included a fair sprinkling of students of economics, biology, and mathematics. In terms of excellence of mathematical background he would rank them as follows: mathematics majors first, followed in order by biology, psychology, education, economics, and sociology majors. Of the education majors one subgroup, those specializing in English, have consistently shown the poorest mathematical equipment, and generally the least "flair" for statistical thinking. This ranking seems congruent with the presumptive interest and cultural antecedents of students majoring in these different fields. Frequently it is helpful to both student and teacher to realize these group differences.

A Background Test, so called because it was used to measure the mathematical background of students entering courses in statistics, was given to 406 students entering, or early in, a course in elementary statistics, drawn from 9 classes at the Universities of Columbia, Harvard, Illinois, Minnesota, Oregon, and Stanford. The courses mentioned were given in psychology and education departments so that if the writer's ranking of the mathematical background of students according to major fields above is fair, the 406 students represent a sampling which is not below average. One form of the Background Test is printed in Appendix A. Students who are desirous of finding out their standing in this fundamental background material may take the test and score it themselves, and interpret it via the percentile norms, given in Appendix A, based on the 406 students.

The idea of making a "good" score on a test has become so established that students will be sorely tempted to cheat themselves. Some of the

ways in which a taker of the test can do it are: *(a)* Read test and scoring key in Appendix A before starting the test. *(b)* When taking the test, not rigidly adhere to time limits. *(c)* Be aware that he is failing because he does not quite understand directions. Look up an answer or two just to get the drift of the thing. *(d)* Before taking the test talk it over with fellow students or instructor. *(e)* When scoring, credit himself with an answer which he intended to be the same as that called correct in the scoring key, but which actually is a little different. *(f)* Generally get rather disgusted with his dumb responses and give a little extra credit here and there. One who does any of these things while taking the test cannot have the satisfaction of an unbiased appraisal of himself in comparison with the run of college students starting a first course in statistics.

A number of very interesting tendencies have come to light as a result of giving the Background Test. It would be appropriate to discuss them here, but this would assist students desirous of taking the test under the standard conditions imposed upon the 406 from whom percentile norms have been drawn up.

Suffice it here to say that the sketchiness of the average entering student's knowledge of arithmetic and elementary algebra would be appalling were it not true that even with such poor equipment students are still able to master some of the broad principles of statistical thinking and treatment of data.

That, with appropriate background, accomplishment would be much greater than now possible has been forcibly voiced by the Committee approved by the Advisory Committee (1933) of Social and Economic Research in Agriculture, making its report to the Social Science Research Council in 1932, entitled *Collegiate Mathematics Needed in*

the Social Sciences. The purpose of the Committee was to determine what mathematics should be taught to students of the social sciences. As the Committee itself pointed out, its report has been largely concerned with the needs of students of economics. There is a slight overemphasis of mathematics connected with temporal series and index numbers in the report of the Committee, as judged by the needs of students of education and psychology, but, with slight modification, their statement is an excellent one as to these needs, and probably even as to the needs of biology students as well. They write:

"The committee believes that a knowledge of the following phases of collegiate mathematics would be quite helpful and useful to a large proportion of the students in economics, and that many students in other social sciences would find it worth while to obtain the knowledge and training that can be acquired through the study of these phases of mathematics, providing it can be done without taking too much time from the study of other subjects.

"1. *Logarithms:* for understanding special plotting methods and forms of equations; for numerical computation in connection with curve fitting and elsewhere.

"2. *Graphs:* (as a mathematical tool) representation of tabulated data on ordinary and logarithmic papers; measurement of slopes and areas (especially in connection with frequency and cumulative curves); graphical determination of maxima and minima; representation of such three-variable relationships as are encountered in statistical studies, by means of three-dimensional diagrams and contours.

"3. *Interpolation:* by reading graphs (ordinary, semi-logarithmic, etc.); by proportional parts; and possibly by successive differences.

. "4. *Equations and forms of important curves:* straight lines; parabolic and hyperbolic curves; exponential and logarithmic curves; logistic curves; sine and cosine curves. The object should be to make the student sufficiently familiar with the forms and characteristics of the simpler curves to enable him to recognize readily the type of curve suitable for fitting to any given body of data, and to appreciate the principal implications involved in the use of such curves.

"5. *Probability:* combinations, binomial theorem, elementary probability; the normal probability curve,—its form, table, and equation; probability and frequency distributions; probability and time series.

"6. *Elements of differential and integral calculus:* the significance of a differential as a limit, as a rate or slope, as a frequency, etc.; partial differentiation; formula for differentiating elementary mathematical functions; procedure for determining maximum and minimum values; integration as the reverse of differentiation; relation between integration and summation; multiple integration.

"7. *Curve fitting* (mathematical principles): plotting tabulated values, their logarithms, reciprocals, etc., to ascertain whether the tabulated points lie on a curve of simple form-determination of the curve through such points, or selected points within the scatter-diagram or chart; the method of least squares or other methods for obtaining the constants of curves of best fit."

The Committee then points out that the background recommended can ordinarily be gotten by pursuing not less than 15 or 20 semester hours of work in a college department of mathematics,

as courses are now ordinarily given, and they observe that this requirement is too heavy a demand upon economics majors, particularly in view of their belief that the desired mathematical training could be given in from 6 to 9 semester hours if mathematics courses were organized with this in view. The Committee recommends the organization of such courses in departments of mathematics. The proposal is an excellent one, but until carried out what is the student of statistics in any one of the social sciences to do? If, in all justice to his program, he cannot take 15 or 20 hours of work in the mathematics department, what is the alternative? That must be decided by each student himself, but the writer recommends that the average student, desirous of mastering the elementary concepts of statistics and not expecting to major in it or to devote his life to research, first, take not less than a five-semester-hour course in college mathematics (a general comprehensive freshman course in college algebra, analytic geometry, etc., if available), and, second, that he do some serious reading upon the mathematical topics just cited to supplement his equipment. As a cue in this the student is particularly advised to remedy his deficiency as revealed in the Background Test of Appendix A. It may be stated that in constructing this test a broad survey was made to determine what are the elementary mathematical concepts demanded for the understanding of current research. Some of the relatively simple mathematical concepts found in the test are not heavily represented in the literature. For example, 11 3/8 divided by 1.65, or $a^3a^6 = a^9$, whereas the more advanced concept $(a + b)^n$, is a common and powerful instrument in statistics. In other words, the student should read with a purpose which is dictated by the needs of the problem of

the subject, not by previous organizations of knowledge. If he wishes to know how to plot y = sine of x, which upon recourse to references he finds discussed on page 300 of some text, it should not be necessary for him to read the preceding 299 pages to secure an answer, but it may readily happen that to understand page 300 he will have to follow back a chain of pages, perhaps like this, 300→299→298→240→241→242→150→84→ a family of concepts known by the student. This type of reading in mathematics is economical of time and effort, and is to be highly recommended. That it can be followed in mathematics may not have occurred to students accustomed to think of the mathematics text as positively sequential, page by page. References could be given to innumerable texts showing where problems of each of the types of those in Appendix A are discussed, but any college student should be able to locate such references himself, and he is expected to do so in connection with each problem that he has missed, if he has taken the test. His arithmetical and algebraic shortcomings, as revealed by the test, especially if so pronounced as to place him in the lower quarter, should be remedied before he starts Chapter VI, and his analytic geometry before starting Chapter IV. This statement asserts that the student is called upon to know certain principles of analytic geometry at an earlier stage than common principles of algebra. This is the case, and it is quite commonly so. The problems of social statistics have not been organized to fit the sequences of the writers of mathematics texts, or the organizers of high school and college curricula in mathematics. We should be content with this situation and expect statistics to provide its own sequences in harmony with its own vista of social problems. An enumeration is here given of certain mathematical concepts which are important for statistics and

which fall under the headings of the report cited:

Logarithms: There are two important sorts, logarithms to the base 10 (here designated "log"), which are used in computational work, and logarithms to the base *e* (here designated "ln") occurring in formulas and curve fitting. The logarithm of a number, say 15, to the base 10, is the power to which 10 must be raised to give 15: thus $10^x = 15$, where x is the required logarithm. The Napierian or natural logarithm is the exponent of a magnitude called e which to the first seven decimal places = 2.7182818. Thus $e^y = 15$, or $2.7182818^y = 15$, where y is the natural logarithm of 15. From the usual table of logarithms we obtain

$\log 15 = 1.176091 = x =$ logarithm to the base 10

Employing a table of natural logarithms we obtain

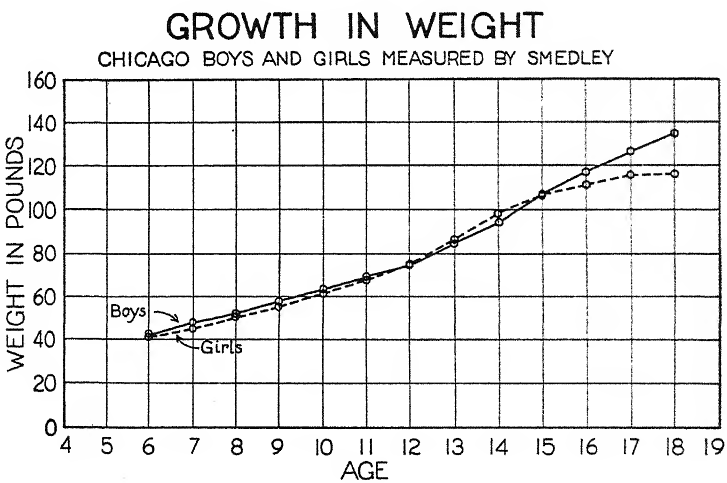
$\ln 15 = 2.708050 = y =$ logarithm to the base e

Either one of these logarithms might readily have been obtained, knowing the other, because of a constant ratio maintaining between them. Logarithms to the base 10 may be gotten from logarithms to the base *e* by multiplication by a quantity M, which to the first nine figures equals .434294482, called the modulus of logarithms to the base 10. Thus $\log a = M \ln a$. For example, 1.176091 = .434294482 (2.708050). Systems of logarithms in addition to these two will scarcely concern the elementary student, but he should be able to use either of these in simple problems. The meaning of terms "base," "modulus," "characteristic" (that part of a logarithm to the base 10 to the left of the decimal point), "mantissa" (that part to the right of the decimal point), "logarithm," "anti-logarithm" (the number corresponding to the logarithm), "co-logarithm"

(the logarithm of the reciprocal of the number) should be known, for there is commonly a peculiar fitness in logarithmic presentation and interpretation of ratios and indexes and of quantities having a natural zero point, as, for example, the price of a commodity or a wage. Where a zero point is not known to exist, or where it is unimportant, or is poorly defined, logarithmic treatment is seldom in order. These brief observations, as well as those of succeeding paragraphs, are not intended to instruct the reader in mathematical background, but to make him aware of the elementary phases of the subject with which he should become acquainted, if not already a master of them.

Graphs: So common a type of curve as that shown in Chart I-I, which the mathematically

CHART I-I



uninformed reader thinks he understands, takes on richness and niceties of meaning when one has a mental picture of the graph of such functions

as the parabola $y = bx^2$, the exponential $y = ae^{bx}$, the Gomperz curve $y = a^{b^x}$, the simple logistic curve $y = \dfrac{a}{1 + be^{-cx}}$ or the growth-senescence curve $y = \dfrac{a}{e^{-bx} + e^{cx}}$. Simple graphic portrayal of data upon two-dimensional paper can commonly be made without any knowledge of the shape and properties of mathematical curves, but an examination of graphs without such knowledge is relatively sterile in suggesting relationships. A knowledge of logistic and more involved curves is not expected of elementary students of statistics, but they should, prior to statistical instruction, be acquainted with such elementary concepts as the straight line and its representation by an equation, the elementary forms of the second degree equation $y = ax^2$, $yx = a$, $x^2 + y^2 = a^2$, the simple exponential $y = a^x$, and they should soon become acquainted with more complex curves.

Interpolation: The veriest tyro is called upon to interpolate when using tabled data. The subject is of increasing importance as more refined work is done. At the initial stage the student should be able to use correctly linear interpolation, that based upon first order differences, in connection with any tabled data. An illustration and definition of certain terms will be made in connection with Table I C. This is the common arrangement of a table of logarithms. The primary use of the table is in getting logarithms, knowing numbers, and its secondary use is getting numbers, knowing logarithms. Based upon its primary use, the means of entering the

TABLE I C

LOGARITHMS TO THE BASE 10

NUMBER	LOGARITHM
1	.0000
2	.3010
3	.4771
4	.6021
5	.6990
6	.7782
7	.8451
8	.9031
9	.9542
10	1.0000

table, called the argument, is a number, and the value gotten out of the table, called the consequent, is a logarithm. The logarithm of 5 is .6990, or, in general terms, we enter with the argument and obtain the consequent. A symbol which will be used in this text to indicate a one-to-one relationship between the measures of a first series and those of a second is $\Bumpeq$, which is to be read "is equivalent to,"—the equivalence being in terms of a given principle of relationship. Thus the statement $5 \Bumpeq .6990$ means that, in terms of the law of relationship given by the table, that is, by the equation $y = \log x$, the magnitude 5, when otherwise expressed, is the magnitude .6990.

The same table may be used to obtain a number corresponding to a logarithm. When so used the entries in the second column become the argument and the entries in the first column the consequent. To find equivalents corresponding to a given magnitude commonly calls for interpolation. Thus, to find the logarithm of 3.5 we go through

this computation:

$$\frac{3.5 - 3.0}{4.0 - 3.0} (.6021 - .4771) + .4771 = .5396$$

or this computation:

$$.5\ (.6021) + .5\ (.4771) = .5396$$

This process is linear interpolation. It is the simplest kind. There is ordinarily an error in the answer obtained. In this case the error is -.0045, because the correct answer gotten from a more detailed table is .5441. We may proceed similarly, as the reader may verify, to find the anti-logarithm of .5540, that is, the number whose logarithm is .5540. The answer is 3.6152 which, from a more detailed table, is found to be in error by .0342.

Arithmetically exact linear interpolation should be within the power of the student before proceeding further, but parabolic and other improved methods involving second and higher order of differences are not expected of him at this time. The reading of a chart or graph with considerable accuracy by means of visual interpolation should also be within his power.

The symbol $\mathrel{=\!\!\circ\!\!=}$ will be found serviceable in many ways, for example, to indicate scores on two tests which are comparable, or equivalent. Thus, if the fifth-grade mean score on the Abb spelling test is 48.7 and on the Baa spelling test is 17.0, these scores represent equivalent degrees of excellence and we may write

$$\text{Abb } 48.7 \mathrel{=\!\!\circ\!\!=} \text{Baa } 17.0$$

It may happen that the consequent for an argument outside the limits of a given table is desired. It can be gotten by a process of extrapolation, but, in general, the process is subject to considerable error. For example, suppose we have given Table I C.

TABLE I C

LOGARITHMS TO THE BASE 10

NUMBER	LOGARITHM
1	.0000
2	.3010
3	.4771
4	.6021
5	.6990
6	.7782

and we desire the logarithm of the number 7. We proceed thus

$$\frac{7.00 - 5.00}{6.00 - 5.00}(.7782 - .6990) + .6990 = .8574$$

This is in error by .0123. If the distance extrapolated is small, the error may not be great, but at best it is more hazardous extrapolating than interpolating, and effort should be made to avoid the necessity of doing so. The more refined methods greatly improve the accuracy of interpolation, but they may not improve extrapolation at all.

Equations and forms of important curves: The student would do well to plot one or more of each of the types listed by the committee and preserve for purposes of reference. The most difficult one to plot is the logistic curve, which calls for the use of logarithms or, most simply, of a "log log" slide rule.

In addition to the curves mentioned, the following are particularly important to the student of education or psychology: the normal curve; the integral of the normal curve, which looks much like the logistic curve, and, along with other curves having the characteristic of a single inflection, is called an ogive curve; peaked, flat-topped, asymetrical, and bi-modal curves, the equations of which are somewhat

involved and beyond the scope of an elementary treatment. Even so, the shapes of these curves and knowledge that concise mathematical statements of each of them are available are valuable facts to know.

Probability: First in this connection is knowledge of the binomial expansions $(a + b)^n$. This is closely associated with the concept of permutations and combinations, and with Poisson's series, point distributions, the normal distributions and with certain skewed distributions. However, most of these topics are appropriate topics for a second course in statistics, so that the beginning student is fairly prepared if he has a knowledge of the binomial theorem and can handle a few simple problems in combinations and permutations.

Differential and integral calculus: An elementary knowledge of calculus is of such vast importance to statistics that beginning students interested in a second or third course in the subject should take an elementary course in calculus, but without such it is to be expected that he would be able to handle a first course in statistics.

Curve fitting and the method of least squares: These topics are equally appropriate to mathematics, engineering, statistics, and certain other fields, but even an elementary knowledge of it is hardly to be expected antecedent to entering a first course in statistics. In fact, it might well be the legitimate subject matter of a second course, and not a prerequisite to it.

Other topics not specifically mentioned in the report of the committee: The propagation of error in the terms of a chain of operations. This issue has to do with the number of figures which are significant in a sum, a product, a quotient, etc., when the number of significant

figures entering into the parts is known. The subject, though fundamental to quantitative study in every scientific field, is seldom treated of in arithmetics and algebras. Accordingly, a brief treatment of the simpler phases of it is given herewith.

In pure mathematics "significant figures" refer to those that are quantitatively meaningful and not merely determinative of the decimal point. Thus in the three following, .7568, 75.68, .007568, the number of significant figures is 4 in each instance. In the case of .007568 (which many scientists would prefer to express 7.568×10^{-3}) the first two zeros merely serve to place the decimal point and are not "significant." To write 756800, so that the reader will know the last two zeros merely place the decimal point and are not themselves significant, it should be given as 7568×10^2 or as 7.568×10^5. If written 756800, it is a figure of six significant figures. Thus in mathematics the number of significant figures in a recorded value is the number of integers recorded, not counting zeros to the immediate right of the decimal point in the case of numbers less than one. This meaning is useful in statistics but another meaning of the term is also important.

When any measurement is taken the number of figures or decimal places to which the measurement is expressed should be limited by the accuracy of the means of measurement employed. If a wooden meter stick is used by a palsied man to measure the length of a live angleworm, and the answer recorded as 7.64381 centimeters, it is obvious that most of the figures put down are not "significant," that is, not trustworthy. Probably the 7 is significant, and the 6 may be better than a guess, but the 4, 3, 8, and 1 are

not entitled to space and attention given them. In fact, they should not appear as a part of the measurement. The measurer may have looked upon the 4, 3, 8, and 1 as representing the best that he could do, but that is not sufficient to warrant recording them. If he felt sure of the 7 and only somewhat doubtful of the 6, it would be appropriate to record the length as 7.6 centimeters. In this case, which is intended to be typical of published data, the last figure recorded is admittedly somewhat in error, but not so much in error as to be no better than a guess, and the figure preceding the last figure is supposed not to be in error at all, except only as there may be a small variation of a tenth carried over from the error in the figure to the right. *Counting decimal places from left to right, we may lay down the rule to record into the figure first thought to be in error by not more than 5.* Thus, if the length be recorded as 7.6, the reader is to understand that the author considers the error in the tenths place to be not greater than 5 in this place, and not less than 1/2, for, if less than 1/2, the recording would have been 7.64, suggesting an error not greater than 5 in the one-hundredths place. Let us write the length thus, $7.\overset{*}{6}4381$, meaning thereby that there is an expected error of not more than 5 nor less than 1/2 in the figure under the star, so that of course the 4, 3, 8, and 1 are meaningless.

So much for original data. Now what happens when numerical magnitudes having errors are added, multiplied, and divided by other magnitudes with errors. Consider

$$7.\overset{*}{6}4381 + 8.4765\overset{*}{3} = 16.12034$$

Over which figure does the star belong in this answer? Consider also the average

$$\frac{7.\overset{*}{6}4381 + 8.4765\overset{*}{3}}{2.00000000\overset{*}{0}} = 8.0617$$

Which figure should be starred in this answer? Clearly the first answer is $16.\overset{*}{1}$, and the second $8.\overset{*}{1}$.

Consider the case

$$\frac{7.\overset{*}{6} + 8.\overset{*}{4}}{2} = 8.0$$

There is an error of unknown amount but presumably not greater than 5 nor less than 1/2 in the tenths place in each addend. To make the problem specific let us say that the error is 1 in this place and that we do not know whether it is positive or negative. Then 7.6 is in error by .1 and the correct value may be written (7.6±.1). Similarly the correct value for the second addend is (8.4±.1), and we have

$$\frac{(7.6\pm.1) + (8.4\pm.1)}{2} = 8.1, \text{ or } 8.0, \text{ or } 8.0, \text{ or } 7.9$$

depending upon which signs are taken. The value 8.0 [given by (7.6 + 8.4) / 2] is thus in error by −.1, or .0, or .0, or .1. The average of these errors algebraically is 0, and the average of them irrespective of sign is .05 [given by (.1 + .0 + .0 + .1) / 4] which is seen to be less than the errors in the measures separately, which was .1. If a positive error in the first addend goes with a positive error in the second, or a negative error in the first with a negative error in the second, then the error in the average is of the same order of magnitude as it is in the addends separately, but if this is not the case, the error in the average is of a less order than in the addends separately.

If the errors in the addends are "random" or

"chance," there is a much less error in the mean than in the measures singly. The specific nature of this decrease in error is shown in Chapter VI where the standard error of the mean is derived. As there proven, *the order or size of an error in a mean is* $\frac{1}{\sqrt{N}}$ $\times$ *the order of size of the errors in the approximately equally reliable addends, where N is the number of addends yielding the mean in question.*

It is left as an exercise for the reader to demonstrate that the error in a difference between two numbers has the same limits as to size as that in their sum. In this instance there is a cumulative effect if a negative error in the minuend tends to be associated with a positive error in the subtrahend, or vice versa.

With this discussion as a background, we may now lay down the principle that the prime consideration in the collection of data to be used in determining means, or sums, or differences, is that they be not subject to "systematic" error, i.e., all prone to errors of the same sign. A very elementary illustration would be the measuring of the height of a certain age-group of boys with shoes on, and reporting the average height for the use of people unaware of the fact that height as reported included shoe heel. We would expect so obvious an error as this to be caught by anyone, but many equally systematic, though not equally obvious, errors enter into economic, psychological, and other nonphysical measurements, ordinarily of great importance to social scientists. *The student of statistics should school himself to be sensitive to possible sources of systematic error.*

Further discussion of this is given in Chapters VI and VII in connection with the standard error

and the probable error of the mean.

If two numbers having errors are multiplied, where is the error in the product? This question may be answered more definitely than that concerned with a sum or an average. Consider

$$7.\overset{*}{6} \times 8.4\overset{*}{8} = 64.448$$

To make the problem concrete and at the same time quite typical, let us consider the error to be 1 in the place starred, but we do not know whether the error is positive or negative. Thus the correct factors and the correct product are given by

$$(7.6\pm.1)(8.48\pm.01) = 65.373 \text{ or } 65.219 \text{ or } 63.675 \text{ or } 63.525$$

whereas the obtained answer is 64.448, so that the error is one of the following: −.925 or −.771 or .773 or .923. The proportionate error in the 7.6 is $\frac{.1}{7.6}$ or .01316. In the 8.48 it is .00118, and in the product it is, regardless of sign, .01435, or .01196, or .01199, or .01432. The average of these, regardless of sign, is .01315, which is approximately the same as the larger proportionate error in the two factors. In fact, we may *in general consider the proportionate error in a product as of the same order of magnitude as that in the factor having the larger proportionate error.*

Consider the two problems

$$31\overset{*}{6} \times .31\overset{*}{5} = 99.540$$

$$31\overset{*}{6} \times .31\overset{*}{7} = 100.172$$

To publish the first product as 99.5 may overstate its accuracy and to publish the second as 100. may understate its accuracy. The precise

number of figures to publish is only accurately determinable by the aid of more information than given and than, in fact, is likely to be available, but as a first approximation we may adopt the rule: *Publish a product to as many significant figures as in the least reliable factor. Publish a quotient to as many significant figures as in the least reliable term.* The number of places carried in computation should be at least one beyond the number to be published, and if a long chain of operations is involved, it may occasionally be wise to carry computations to two places beyond that demanded for publication.

It is suggested that the reader investigate by cut and try methods what modification or refinement of this statement is needed to make it more accurate. As examples, when the proportionate errors are the same in the two factors, consider at least two cases: (a) when the error is not in the first significant figure, and (b) when it is.

The novice in experimental investigation, combining three statistics such as the following

$$10\overset{*}{3} \times .03\overset{*}{4} \times 1.20\overset{*}{0} = 4.2024$$

to obtain a final outcome, is likely to be uncritical in the distribution of time and effort which he puts upon the various aspects of his problem. Suppose the terminal statistics equals $4.\overset{*}{2}024$. Suppose .034 is gotten by using an instrument, for example a psychological test, provided by some earlier worker, and 103 and 1200 by using newly devised instruments. Refinement in these, leading to measures $10\overset{*}{3}$ and $120\overset{*}{0}$, has been accomplished at great labor and it is futile in view of the error inherent in $.03\overset{*}{4}$. That a

chain is as strong as its weakest link has been overlooked time and again by semi-competent statisticians.

On the other hand, in many statistical problems analogy with a chain is unsound. Thus, characteristically, the mean is not as unreliable as the most unreliable measure which has entered it. Here the analogy may be to a rope of many strands, whose strength is much greater than its weakest strand.

It is sound practice so to report quantitive results that a figure based upon observations such as, say, 1.743, carries with it the meaning that the 3, but not the 4, may be expected to be in error. Suppose a finally computed statistic is 174328, with an expected error in the last three figures. Then the result should be reported in some such manner as 1.743×10^5. The mere fact that the decimal point happens to be one or more places to the right of the last significant figure does not constitute warrent for publishing the meaningless figures, such as are the 2 and 8. An investigator should think of a computed statistic as in three parts, thus: 174|3|28, the first nearly indubitably trustworthy, the third substantially meaningless, and the middle part, of one digit only, of questionable significance, and he should publish the first two parts only.

In a formula involving several terms, some of which are much less reliable than others, there is generally little point in keeping the more reliable figures to as many places as they are significant, for the error in the outcome is commonly determined by the error in the factor having the largest error. When addition and subtraction are involved, the largest absolute error in the parts is to be considered, and when multiplication and division are involved the largest proportionate error is the error of

importance. A thorough discussion of this matter is to be found in Scarborough's *Numerical Mathematical Analysis*, (1930) and an excellent discussion, with numerical illustrations, is given in Walker's *Mathematics Essential for Elementary Statistics*, (1934).

And finally, as an essential part of the background of the student of statistics, should be mentioned *computational accuracy and knowledge of methods of checking the same.* Many a student lacks computational accuracy in simple arithmetical processes, and few have the habit of checking a computation by performing it in a second and independent way. The idea that the need for such a habit has passed with the advent of computing machines is stultifying. The writer has found no operation in greater need of checking by the ordinary student than that of extraction of square root. The habit of proving, by squaring, the root found is simple and very effective.

We may assert that there is need for a proper psychological background. One's processes in arithmetic and algebra have commonly been built up through the vehicle of abstract numbers. Many of the magnitudes entering into a statistical study are not of this sort, while others are.

For example, in $y = 3a^2x^b$ it might be that 3, 2, b are abstract quantities, not subject to redetermination or experimental change, operating upon the other measures and enabling them to be expressed in an equivalent form y. If so, the abstract number concept with reference to these is appropriate. The magnitude a might be a statistical concept, earlier determined by oneself or by another investigator, and not intended to be changed in this experiment, but nevertheless an experimental constant and not of the nature of a pure abstract number. The magnitude x may be the crucial observation of this experiment.

As it is operated upon, every derivative from it retains the stigma of its origin. If the observation x has a meaning and if the final function of it y has a meaning, then every intermediate function has a definite physical meaning, and the proper psychological attitude of the student is to trace this out step by step. It is questionable whether it ever happens that the meaningfulness of the outcome is accurately and fully appraised if the meaning of the successive steps in the process has been obscure. Certain magnitudes that the statistician is working with are living and vibrant in a sense that the abstract symbols are not. The 3, 2, b, and a provide the form of expression but the life that flows through is the x, and it is this life that gives primary meaning to each successive function that is dealt with. It is a social blood stream flowing through algebraic arteries.

Plan of study: The table of contents provides the plan of this text. The first ten chapters constitute an integrated sequence built upon the background as indicated in the preceding paragraphs. The later chapters provide occasional advanced techniques. In general, in these chapters, citations to sources supplant full derivations. This text may be used as a basis for an advanced course in statistics, especially by pursuing the references and developing the techniques of the later chapters. The student studying it as a first course should realize that the early chapters give the first steps in many sequences. Following upon elementary graphic work of this text is graphic analysis; upon a study of means and standard deviations is a study of other descriptive features of a distribution; upon simple correlation of quantitative measures is correlation between two variables expressed in categorical and ordered series, and correlation issues involving more than two variables; following the

normal distribution in one variable is the normal correlation surface; following linear regression, which is equivalent to fitting a straight line, is curve fitting of all sorts; and paralleling an algebraic interpretation, running throughout all the preceding, is a geometric and trigonometric interpretation involving more than three dimensions. A member of an elementary class in statistics made a most stupid remark, unless perchance it was a bit of wry humor. He said, "The work of this course gives practically all the statistics that a schoolman could ever use, doesn't it?" He might as well have observed that only little care in thinking is useful if one is engaged in a small business.

An early step in a statistical problem is the collection of data. This problem has been treated of in general terms (see Keynes, 1921) and, of course, in detail in connection with many special problems, particularly those of economics (see Zizek, 1913 and Young, 1925). Because of its almost infinite ramifications, depending upon the purpose and amenability to control of the phenomena under observation, the process is not seriously presented in this text. Its study is most appropriate after the purpose of an investigation is clearly defined, and after physical limitations of time and expense are known.

A number of terms have been used in this chapter which have technical statistical or mathematical meaning. Since this section is concerned with the appropriate background for a study of statistics, these, together with certain other equally elementary statistical terms and symbols, are defined and criticized herewith. Additional terms will be similarly presented in subsequent chapters. The arrangement is alphabetic except for such closely related terms as call for definition together. The explanations of terms here given are intended to be accurate as far as they

go, but not necessarily complete because it is thought that definitions so rigorous as to include the fields of meaning of higher mathematics would be confusing to the elementary student.

absolute value: the value of a magnitude taken positively, thus the absolute value of -7 is 7, and the absolute value of x is $(-x)$, if x itself is negative. Bars enclosing a quantity indicate absolute value, thus: $|x|$, or $|-7|$.

abscissa: see *coordinates*.

arbitrary origin: see *coordinates*.

argument: the value with which a table is entered when a related value, called the *consequent*, or tabled value, is sought. If one finds, from a table, the logarithm of 5, the number 5 is the argument, and the logarithm, namely .69897, is the consequent.

axis: a straight line in a figure with reference to which the figure is symmetrical, or one of the axes of reference. See *coordinates*.

biometry: the application of measurement to living things.

correlation: the tendency of paired measures to vary concomitantly. This term is further defined in Chapter X.

crude score: see *score*.

consequent: see *argument*.

constant: see *literal notation*.

coordinates: there are two kinds commonly used in connection with two-dimensional representation, rectangular and polar. Rectangular coordinates consist of two lines at right angles, the horizontal axis, or *abscissa*, and the vertical axis, or *ordinate*. The intersection of the axes is the *origin*, or zero point, from which measurement is taken. Two variables related by some

equation or law may be represented, one by distance in the direction of the *abscissa*, and the other by distance in the direction of the *ordinate.* Paired values define a single point on the two-dimensional surface. The connection of all such points yields a curve (still called a curve though a straight line or composed of straight-line segments). If this curve is such that, for each value of the first variable, there is one and only one value of the second, the first variable is a *single valued function* of the second. If at the same time the second is a single valued function of the first, then the curve is a *monotonic* curve,—one that is always rising or always falling. The singular "coordinate" refers to distance in either one of the two dimensions. The plural "coordinates" refers either to the two coordinates for some point or to the two lines at right angles through the origin, which are also called the axes of reference. If the point from which measurements are taken is chosen for convenience, or if it has no special physical significance, the origin is called an *arbitrary origin*. *Polar coordinates* define a point in a space of two dimensions by distance from an origin and angular distance from a given line. This system is less commonly employed in elementary statistical work than rectangular coordinates.

curve: see *coordinates*.

degree: in trigonometry, one three-hundred-and-sixtieth of the circumference. The degree of an equation is the highest sum of the exponents of any term of an equation, thus, $4x + y = 7$ is of the first degree, $4x^2 + y = 7$, or $4x + xy = 7$, of the second degree, etc. See *dimension*.

dimension: a dimension of an equation or of a term of an equation is the same as the degree of it. Thus xy^3z^2 is of the sixth degree or sixth

dimension. See *degree*.

distance: commonly used to indicate not merely spatial dimension, but also dimension in any quantitative function. Thus we may speak of the distance one score is from another, though the function may be intelligence, reaction time, or whatnot.

equivalent: equivalent measures are those expressing the same thought in different terms, thus 39.37 inches are equivalent to 100 centimeters. See page 32.

evaluation: see *formulation*.

factor: in addition to the ordinary meaning that a factor is one of the quantities entering into a product, the term is used quite extensively to indicate one of the components entering into a sum, particularly if the components making up the sum are uncorrelated with each other.

function: this is an aristocrat of mathematical terms. It is indifferent to such lesser terms as variable, constant, exponent, such operations as multiplication or addition, and such specific relationships as subordination, correlation, association, though all of these serve it and give it specific meaning. To say that y is a function of x asserts that operating in some manner, which he who knows may designate, upon x yields y, and furthermore that y is gotten in no other way. Distinctions which are commonly important for statistics between functions may be found in whether the function is *multiple valued*, *single valued*, or *monotonic*, as explained in a preceding paragraph in connection with curves plotted upon rectangular coordinates.

formulation: the determination of the specific nature of a function. *Evaluation* is the process of finding the numerical value of y which is a function of x, for a given value of x.

imaginary number: in mathematics the square root of a negative quantity is called an imaginary number, and is contrasted with real numbers, those having finite values between $-\infty$ and $+\infty$. In addition to quantities imaginary in this respect, there are in statistics values which are "impossible" or "unreal", such as a product-moment correlation coefficient greater than one, a reliability coefficient less than zero, or a negative standard deviation. To avoid confusion, these latter would better not be designated imaginary, however much they are a matter of fiction.

increment: a term used ordinarily to indicate a small value or a small change in value and quite commonly represented by δ or Δ. Such a small increment approaches the differentials of calculus as it becomes smaller and smaller.

invariant: this highly useful concept, wherever employed, is peculiarly valuable in connection with mathematical and statistical problems. If the vertical, the north-south, and the east-west dimensions of an egg are taken as it is held in various positions, there result a great many values, all different from each other. These dimensions are variable as *transformations*,—changes in position,—are made. (This geometric explanation of "transformation" is in harmony with the algebraic explanation given under the term *transformation*.) If the ratio of the longest diameter to the shortest diameter is computed for each position, it remains the same and is accordingly invariant under the transformations. These transformations may be considered to be merely different ways of viewing the same thing. If a thing is chameleonlike and looks different as each point of view is taken, there is no virtue in it, but if some aspect of it can be selected such that it is the same no matter the point of view (in mathematical terms,

100161

no matter the axes of reference), it becomes a thing in which there is no shadow of turning, and in which truth resides. The discovery of the invariants in phenomena is the main concern of statistical and experimental science.

literal notation: three sorts of elements enter into algebraic expression: *(a)* symbols of operation, or of relationship, *(b)* symbols indicating invariable features of the equation, that is, *constants*, and *(c)* symbols indicating variable features, that is, *variables*. A large class falling under *(b)* consists of the ordinary cardinal numbers. With quite a number of exceptions, it may be said that in this text letters near the end of the alphabet, particularly x and y, will represent variable quantities, letters near the beginning of the alphabet, particularly a and b, will represent constants, or invariable quantities, and symbols and Greek letters will represent operations. For more detail see the paragraph following upon *symbols*.

maximum and *minimum*: the meaning of the terms will be obvious to the reader, but their importance in statistical work is so great that they should receive definite obeisance whenever met in a statistical study. To assert that one's purpose is to get the maximum of enjoyment out of life is a literary pleasantry, but to say that the object is to make errors minimal should hold the attention and vivify the interest until the means for accomplishing this become clear. Do not slip over these key words to understanding, these cues warning one of a chance to observe a bit of intellectual cleverness. The neatness of mathematics is incorporated into statistics through these concepts.

monotonic: see *coordinates*.

observed value: this describes the value of a variable in distinction *(a)* from the value esti-

mated from a knowledge of one or more correlated values, or *(b)* from an unknown true value.

origin: see *coordinates*.

orthogonal: see *space*.

percentage: a percentage is 100 times an equivalent proportion. If 5 out of 20 belong in a certain group, the *proportion* in this class is .25, and the percentage is 25. The subtlety of this relationship is not so great as to provide an excuse for the common misuse of these terms.

proportion: see *percentage*.

real number: see *imaginary number*.

score: a quantitative or qualitative recording of an observation.

space: in addition to the familiar variety, the student of statistics is frequently concerned with an *n*-space, or an *n*-dimensional space, where *n* is greater than 3, because he can so readily represent four or more variables each in an independent direction (*orthogonal* or at right-angles to each other) in this space. The mental picture of four lines through a single origin, each at right angles to the other, is beyond one, and the student is not advised to weary his intellect with trying to make such a picture. Rather he should manipulate his four or more variables according to the rules of algebra, or he should reason by analogy with the geometry of two or three dimensions. In other words, when an *n*-dimensional proposition is presented do not become panic-stricken and consign the treatment of it to mortals cast in a higher mold and endowed with a pineal eye.

statistic: a statistic is any quantitative or qualitative characterization of a thing observed, or any summarized statement of such. *John Doe, aged 12, height 59 inches*, is one of a *group of*

70 whose *mean age is 13.0, average height 63.3 inches,* in spite of *one member being but 48 inches tall.* Every item in italics, as well as an innumerable number of others not mentioned, is a statistic. *Statistics* is the plural of statistic, as defined, and also the entire body of sound practice which has been developed for summarizing observations and discovering invariant features of them. These summarized statements, such as *mean age 13.0,* do not "prove" points. They merely describe a situation. The expression "statistics prove" is always a misstatement. Statistics are their best "measure," permitting him who will to "infer." We correctly say that a yardstick "measures" the height of a child, not that it "proves" it.

symbols: it is thought that a mere mention of most of the symbols to be employed in this text should suffice, so that explanations of only a few follow:

$=$ equality

$\equiv$ identity

$\doteq$ or $\div$ approximate equality

$\asymp$ equivalence, see page 34

$+$ plus

$-$ minus

$\div$, —, /, : division, thus $a \div b$, $\frac{a}{b}$, a/b, $a:b$

$\times$, $\cdot$, or nothing at all, meaning multiplication, thus, $a \times b$, or $a \cdot b$, or ab

∞ or ω an infinitely large magnitude, or a "true" magnitude, i.e., one having no chance error in it.

exponent, to indicate a power, thus

$$a^5 = a \cdot a \cdot a \cdot a \cdot a$$

$\sqrt{}$ square root, $\sqrt[3]{}$ indicates cube root, etc.

fractional exponent, to indicate root, thus

$$a^{\frac{1}{5}} = \sqrt[5]{a}$$

subscript, to specify the one of several of a sort, thus x_1, x_2, ordinarily stand for score or measure on a first variable and on a second. x_{1a}, x_{1b}, etc., to indicate score of subject *a* on the first variable, of subject *b* on the first variable, etc.

Σ, Greek capital sigma, to indicate the sum of all the measures in question, thus Σx_1 equals the sum of all the x_1 measures for the group in question, equals

$$x_{1a} + x_{1b} +, \cdots x_{1n}$$

if there are *n* measures.

... to indicate omitted item of a series, as shown just above.

)(to indicate excluded term of a series, thus,

$$r_{12.3 \cdots)7(\cdots 9} = r_{12.345689}$$

Π Greek letter pi, used as a symbol of operation indicating the product of all the measures in question, thus,

$$\pi\, x_1 = x_{1a} \cdot x_{1b} \cdot x_{1c} \cdots x_{1n}$$

if there are *n* measures.

e = 2.71828183, the Naperian base of logarithms, which is a magnitude entering into many formulas.

N indicates the number of measures in a series. When used as a subscript x_{1a}, x_{1b}, $\cdots$ x_{1n}, *n* may be used in lieu of *N* because its size makes it better adapted for use as a

subscript.

transformation: the substitution of a second variable for a first of which it is a function. generally a linear function. Thus, if relationships involving x are given, and in lieu thereof relationships involving y are determined, where $x = a + by$, a linear transformation has been made. See *invariant*.

transmutation: used synonymously with *transformation*, which is the mathematical and approved term. No alchemical implications are involved.

variable: see *literal notation*.

SECTION 5. PLAN OF STUDY

If the reader has looked upon statistical method as a tool to be resorted to when necessity dictates, it probably holds a very lowly status in his mind. In one sense this is correct, for statistics is definitely a servant to the sociological, biological, and politcal sciences, but a very independent servant in his methods, a servant who cannot be ordered to "prove a point," but only to investigate it; a servant who repeatedly tells his master, the setter of the problem, that he is on the wrong track, that the problem is incapable of solution with the data at hand, or that refinements of procedure called for if a solution is to be obtained are far beyond those anticipated.

As an adviser of graduate students it has been the writer's experience to have "worthwhile" issues presented, and time and again with no, little, or inadequate knowledge of what is implied in reaching a solution. As one of many examples he could cite the case of the student who wished to prepare himself to make analytical diagnoses of abilities of pupils and to give them sound educational and vocational guidance, but who had no slightest idea of studying or utilizing partial

or multiple correlation. To such a student statistics asserts its independence, demands techniques, defines the limits of undertakings as dependent upon data and treatment, and finally pontificates in the matter of reliability of diagnoses. Statistics is the servant in the house, but, whether known or unknown to the master, is guided by rules of conduct so full of virtue that they may not be overstepped.

The student who really attains a frame of mind that is sympathetic to statistics must find satisfaction in neatness, dispatch, soundness, and generality of procedure, not satisfaction only in the proof of a cherished belief. In fact, the first-mentioned satisfactions should be so great that a disproof of a belief carries but a minor sting.

The usual elementary text in statistics seeks to acquaint students with the techniques necessary to obtain a number of highly valuable summarizing statistics: averages, measures of variability, a few measures of correlation, and mastery of some of the elementary graphic devices. This text considers these things as a secondary aim, the primary aim being to acquaint the student with the logic back of these procedures. It is believed to be more important for a student to understand why a mean is employed than to know the steps for calculating it; why and when a product-moment correlation coefficient is significant than to master the intricacies of the xlnt correlation chart. This broader purpose necessitates a different and a more extended treatment of so-called elementary issues than is usual. It is accordingly here attempted to cover in the earlier chapters a smaller number of different statistical concepts than are usually covered in an elementary book, but to do so with greater than usual logical completeness.

In that this text is designed to serve as an

introduction to advanced study, the notation used is that appropriate to advanced statistics. For example, it would be quite possible to write an elementary statistics without recourse to letters in the Greek alphabet. This, however, is not here attempted, for Greek letters will not only serve in the elementary problem but in addition will provide familiarity with concepts essential in more advanced statistics. Special care has been taken to provide the simplest notation congruent with prospective needs of the student. This occasionally, but only occasionally, leads to a notation which is a trifle more elaborate than the immediate problem demands.

PROBLEMS

Problem 1.

As a vocabulary review the student may test his understanding of the following words and phrases.

abscissa	distance
absolute value	equality
arbitrary origin	equivalent
argument	evaluate
axis	factor
biometry	finite
central tendency	formulation
consequent	function
constant	identical
coordinate	identity
coordinates	imaginary
correlation	increment
crude score	indeterminate
curve	interpolation
degree	invariable
determinate	invariant
diameter	line
dimension	mathematics

maximum
minimum
monotonic
multiple valued function
negative
ordinate
origin
origin, arbitrary
polar coordinates
positive
real number
reduce
score
single value function
solution
space
statistic
sum

Problem 2.

With the general principles covering the propagation of error in mind, place a star over the appropriate figure in the right-hand members of the following equations to indicate the most probable location of the errors in the answer consequent to errors as designated by stars in the terms of the left-hand members.

1. $123.\overset{*}{5} + 11.6\overset{*}{2} = 135.12$
2. $154.4\overset{*}{2} + .01\overset{*}{4} = 154.434$
3. $95.7\overset{*}{5} - 13.\overset{*}{4} = 82.35$
4. $67.80\overset{*}{5} - 13.6\overset{*}{0} = 54.205$
5. $16.1\overset{*}{0} + 5.30\overset{*}{0} = 21.400$
6. $10\overset{*}{3} \times .30\overset{*}{5} = 31.415$
7. $41.\overset{*}{5} \times .01\overset{*}{4} = .5810$
8. $.0\overset{*}{7} \times 8.\overset{*}{6} = .602$
9. $2\overset{*}{0} \times 15.\overset{*}{0} = 300.0$
10. $.00\overset{*}{7} \times .4\overset{*}{4} = .00308$
11. $108.\overset{*}{6} \div 3.\overset{*}{0} = 36.2$
12. $.074\overset{*}{7} \div .124\overset{*}{5} = .6000$
13. $19\overset{*}{2} \div 1.\overset{*}{6} = 120.0$
14. $27\overset{*}{3} \div .00\overset{*}{7} = 39{,}000$
15. $(2.\overset{*}{7})^2 = 7.29$
16. $(1.\overset{*}{5})^3 = 3.375$
17. $(.001\overset{*}{2})^2 = .00000144$
18. $\sqrt{22\overset{*}{5}} = 15.00$
19. $\sqrt{3.6\overset{*}{1}} = 1.9000$
20. $\sqrt{.00052\overset{*}{9}} = .02300$

These exercises have been given because of

the common error of recording far more figures than are significant.

The following problems in plotting mathematical equations range from very simple to difficult, though these latter (Problems 7 and 8) involve nothing beyond the handling of logarithms and negative exponents.

Problem 3. Plot the straight lines

(a) $y = 4 + 3x$

(b) $2y = 4 + 3x$

Relate the slope of each line as shown graphically with the coefficients of x and y.

Problem 4. Plot the second degree parabola

$y = 1 + 3x^2 + x^3$, which may be written

$y-5 = (x-1)(x+2)^2$

Relate the graph to this second way of writing the equation.

Problem 5. In Table XV C the proportion of cases in a unit normal distribution lying to the left of the corresponding x-value is p. The equation relating p and x is

$$p = \int_{-\infty}^{x} z\,dx, \text{ in which } z = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

Since the related values of p, z, and x are given in Table XV C no computations are necessary in making the following graphs (a) z as a function of x, and (b) p as a function of x. (a) is the highly important normal distribution and (b) is the graph of its integral.

Problem 7. The simple logistic is an important growth curve. Its equation in the general

form is

$$y = \frac{k}{1 + e^{a+bx}} \quad \text{or} \quad y = \frac{k}{1 + Ae^{bx}}$$

Dr. Marian Wilder has determined the constants k, a, and b, for the logistic which fits as closely as possible (in the least squares sense) to the integral of the unit normal distribution. The equation is

$$y = \frac{1}{1 + e^{-1.6637x}}$$

Plot this and compare with *(b)* of Problem 6.

Problem 8. The Gompertz curve

$$y = a^{-b^{-x}}$$

is also a growth curve. Dr. Marian Wilder has determined the constants a and b for the Gompertz curve, which fits as closely as possible (in the least squares sense) to the integral of the unit normal distribution. The equation is

$$y = 1.9019^{-3.1500^{-x}}$$

Plot this and compare with *(b)* of Problem 6.

CHAPTER II

STATISTICAL SERIES

SECTION 1. TYPES OF DATA

Logically we may classify series as follows:

I. Temporal: *(a)* qualitative or *(b)* quantitative differences, shown as time changes.

II. Spatial or Geographic: *(a)* qualitative or *(b)* quantitative differences shown as location in space changes.

III. Momentary (or Pseudo-momentary): *(a)* qualitative or *(b)* quantitive differences shown as the items in a sample change.

Under pseudo-momentary fall the large class of observations which it is assumed have not been altered by the particular moment or place of observation. For example, frequently the records of a week ago, a year ago, or of today, of a chemist may all be thrown into a common distribution, believing that the moment or place of observation has been immaterial to the issue at hand.

The following naming of series has been here adopted:

I. *(a)* Qualitative-temporal series.
(b) Temporal series.

II. *(a)* Qualitative-spatial, or qualitative-geographical series.
(b) Spatial, or, if related to location upon the earth, geographical series.

III. *(a)* Qualitative series.
(b) Quantitative series.

Let us consider these in more detail.

Temporal series: If a single thing, say a child, is observed at successive periods of time with reference to a single character (the term preferred by biologists), or trait, or characteristic, a time series results. Usually the successive measures are quantitative, as would be the case if height is measured, but they need not be. Thus if the successive observations of an infant's vocalization give the syllables blah, ma, bab, omp, these constitute a qualitative time series, as the differences in the trait, as time varies, has been the issue.

Table II A is another illustration of a qualitative-temporal series. As is generally the case, it is here obvious that fuller information would disclose quantitative phenomena underlying most of the qualitative statements listed.

TABLE II A

IMPORTANT STEPS IN THE AUTOMOBILE INDUSTRY*

1900	"Mass production" of motor cars first employed
1901	First American speedometer made
1902	Alloy steels used in making automobiles
1904	Head lamps included as standard equipment
1906	Front bumpers and electric horns pioneered
1908	Left-hand drive popular
1909	First closed bodies built

* AUTOMOBILE FACTS, II, No. 2, October, 1939.

TABLE II A

(CONTINUED)

1911	Electric starting, lighting, and ignition first combined in a single system with storage battery
1912	Windshields made as part of body by one company
1915	Top and windshield became standard equipment
1916	Steel frame body introduced
1917	Disc wheels appear
1918	Tire sizes standardized
1919	High-pressure chassis lubrication adopted
1920	Windshield wiper universal
1921	Hydraulic brakes on some cars
1922	Balloon tires appear
1924	Bumpers became standard equipment
1925	Four-wheel hydraulic brakes and all-steel bodies used
1926	Rubber engine mountings, rubber spring shackles, and adjustable front seats introduced
1927	Gearshift standard on all cars
1928	Safety glass introduced as standard equipment
1929	Automobile radio introduced
1931	Vacuum spark control used
1932	Drop-center rims replaced "detachable"
1933	First steel top used on some cars
1934	Synchronized front and rear springs introduced
1935	All steel bodies universal. Steering-post gearshift on one make
1937	Built-in windshield defroster ducts
1938	Improved windshield vision, independently sprung front wheels attract attention
1939	Steering-column gearshift becomes popular
1940	Sealed-beam headlights universally approved

If one child at age 18 months is measured for some trait, a second child at age 24, and a third at age 39 months, etc., there results a series (a pseudo-time series) which is not a true time series, for no single principle of sampling has

been employed, that is, the successive items represent differences with reference to *(a)* time and *(b)* subject. When such series are treated as time series it is upon the assumption that differences in *(b)* are immaterial. This is generally a hazardous assumption and should call for specific study before being made. Most achievement and intelligence test norms which are intended to be time series or growth curves in the function in question involve this assumption, for it has seldom been attempted to base norms upon the same subjects as they age from year to year. It is commonly, for example, assumed that the thirteen-year-olds fairly represent what the twelve-year-olds will become in a year's time. Expediency is the excuse for this method of approximating true time and true growth curves.

If the trait measured at successive periods of time is a developing one, it may be called a growth series, and, when plotted, a growth curve. A growth curve generally starts at a point near zero. It is not to be expected that the exact starting point can be caught; even if the start in the case of a living bisexual organism is called the moment when the sperm impregnates the ovum, still the point is arbitrary, for both sperm and ovum have had antecedent histories. Starting with an initial measure, near zero, the typical growth curve increases, with first increasing and later decreasing acceleration, to some physiological maximum at a date which may be fully as difficult to determine as the zero point date.

In contrast to this typical growth curve is the time series characterizing a much later stage. An economist may not know the time and correlative price of the first bushel of wheat. The starting point is lost in the remote past, but the fluctuation of the price of wheat from day to day and month to month, as time is now changing, yields

a very typical time series, but it is no longer called a growth series.

Typical issues arising in connection with time series are not the same as for other series. In Chapter V it is pointed out that the essential issues, the essential techniques, and the essential summarizing statistics, all change as type of series changes.

As a second type of series may be mentioned the geographic series, or, to use the more comprehensive term, the spatial series. In this case differences in time are either (a) avoided by collecting all items at the same time, or (b) considered irrelevant, but locations from which the data are collected are of prime importance. If the material collected has been distributed in three-dimensional space, the very general spatial series results. An illustration of such a series would be a survey of the dust content of the air in the different city precincts and at different elevations above street level. Problems involving data of this three-dimensional sort are becoming more common because of air and submarine transport. A not too large portion of the surface of the globe may be considered a plane, so that the ordinary spatial series involves differences in locus of items in two dimensions only. The series is then called a geographical series. Each term collected is uniquely and intrinsically associated with the point of origin. If this fact is lost in any step of procedure, the series immediately ceases to be a geographical series. If in a plot of ground laid off like a checkerboard, one unit of area reveals 2 cut worms, a second unit 23 ants, a third 1 glow worm, etc., these successive qualitatively different items attaching to the differences in location constitute a geographical series. The more ordinary sort is found when the differences revealed are quantitative, as would be the case if one unit of area, perhaps an entire

township, yielded 10,000 bushels of wheat, a second zero bushels, a third 5,000 bushels, etc. The essential statistics of geographical series are even more specialized than are those of a time series.

The map is essential in portraying spatial series. Many spatial series show both qualitative and quantitative differences, in which case a considerable ingenuity is needed to devise a map with cross sectioning, or color scheme, to portray the essential facts. Spatial series are intrinsically more amenable to graphic treatment, and less to numerical treatment, than temporal or quantitative series. The maps of the U. S. Coast and Geodetic Survey, of the Weather Bureau, and of the Census Bureau show the high degree of completeness, variety, and detail of portrayal that is possible The groupings of territories in spatial series and the subdivision of areas may follow conventional procedure or the peculiar needs of the problem. The order adopted by the Census Bureau in giving population statistics is shown in Chapter IV, Table S.

The third and fourth types of series are the qualitative and the quantitative which exist when the principle of sampling has either allowed for time and space by making them constant, or when it is considered that such differences in time and space, at which the successive observations have been made, are immaterial. Thus some important consideration other than, or in addition to, time or space has been utilized as the principle in getting the cases of the sample, and of course some definite principle has been followed in making the observations or taking the measurements of the case. If the resulting characterizations of the observations taken differ qualitatively one from another, the series is a qualitative or categorical series, whereas if they differ in terms of amount it is a quantitative series. Observations upon

eye color, or hair color, or blood type, or sex, or abnormality of fifth-grade school children, would yield qualitative series, whereas recording the height, or weight, or age, or psychological test score, or achievement test score, etc., would yield quantitative series. One immediately deducible property of the items of a quantitative or qualitative series is that they are not tied to specific time and place of origin, that is, these are neglected when studying them.

Thus in every instance and for every sort of series the observations taken cover but a fraction of reality. If numbers of bushels of wheat are the elementary statistics, the animal life in the unit of area is not. If heights of twelve-year-old boys are the elementary statistics, their places of birth and dates of birth, except as such might define the principle of sampling twelve-year-olds, are not. And so it goes. We are continually studying aspects of total situations, and it must be so, whether the method of study is admittedly statistical or not.

When a research undertaking is planned, it will be found conducive to later precise treatment if an attempt is made to relate the issues to be studied to some one of the four types of series. If this is impossible, the statistical handling of the resulting data becomes complex, though not necessarily insoluble.

The classification here followed of simple series as being temporal, geographical (or, more generally, spatial), quantitative, and qualitative, is not identical with that made by certain statisticians: Yule (1929) divides statistics into the statistics of attributes and the statistics of variables. These correspond to qualitative and quantitative, as herein given. Jerome (1924) divides series into historical and cross-section, and the latter he divides into geo-

graphical, qualitative, and quantitative, thus agreeing with the classification herein used. Another common classification is into discrete and continuous series. This is made by Garrett (1926), Thurstone (1925), Rietz (1924), and Young (1925). R. A. Fisher (1934, page 25) classifies certain series as temporal and again he classifies them (page 4) as continuous and discrete. It should be noticed that quantitative and qualitative are not synonymous with continuous and discrete. The initial data in quantitative series may be continuous, as, for example, heights of twelve-year-olds, or discrete, as, for example, the numbers of children in families; but the initial data in qualitative series can only be discrete. The merit of any classification is in its utility. In general, the reader will find that a definite set of statistical techniques follows from the type of series involved. This is true for the four types here employed. If the quantitative is further subdivided into continuous and discrete, it will be found that there are separate statistical techniques which are in harmony with each.

SECTION 2. TYPES OF ISSUES

We have considered the types of data occurring in the different series. Let us now inquire as to the differences in issues consequent to differences in series types.

In time series one is concerned with fluctuation of some single function with changes in time. If the function is greater at one time than another some measure of this excess is important. If a natural zero point, or upper limit, such as a 100 per cent point, is not available, then some gross measure of difference must be employed, but in the very common case where one or both of these natural limits are known (limits in terms of the amount of the magnitude, not

limits in terms of the time at which the zero amount or 100 per cent amount is attained), the obvious measure of relationship is the ratio of the magnitude of the function at one time of its magnitude at another, and since the zero magnitude at the beginning of growth is never an observed value, one does not secure infinite ratios. Should one secure an infinite ratio because the magnitude at some time after the start becomes zero, then the ratio as a statistic loses its value. *In general, the ratio is a statistic peculiarly adapted to the study of temporal series.*

The price of a commodity at a given date, divided by the price at a second date, is such a ratio. If a second commodity for the same two dates is involved, a second price ratio is obtained. *Composites and aggregates of such ratios* or relatives *are called* (price) *indexes, and these are essential tools for the study of time series.*

If the price in 1934 is expressed relative to the price in 1933, the year 1933 is called the basal year. The choice of such basal moments of time with reference to the magnitude at which time all other magnitudes are expressed is essentially a problem of time series.

There is frequently evidence of a systematic change in magnitude over a considerable period of time. This is called a trend, and a line indicating it when the series is plotted is a trend line. The trend is an essential statistic of many a time series.

There may be a regular order of fluctuation as time changes, yielding cyclical phenomena, and a study of these and their periods are peculiar to time series. Any time series of appreciable duration (in studying etheric vibration .001 of a second would be a very appreciable duration) may be expected to show periodic fluctuations.

As a consequence one of two procedures is necessary, dependent upon whether it is desired *(a)* to study the changes within a certain cyclical period, or *(b)* to study trends independent of such periodic changes. Illustrations will make the problem clear:

(a) Let it be required to ascertain the nature of the load of an electric-power generating plant during a twenty-four-hour period. The current consumed per hour for some one day could be tabulated or plotted. The result would have only such accuracy as would result from a single day's sampling. To obtain a more reliable picture, a number of days could be combined and the tabulation made showing the average load for each hour of the 24. Obviously error might creep in here, for the load on a Monday would be quite different from that on a Saturday or Sunday, and perhaps different from that on the other days of the week. With due allowance for holidays, probably a very satisfactory idea of the hourly fluctuations of the Monday load could be obtained by pooling results for several Mondays. Differences in daylight, temperature, etc., would make it unsound to combine all the Mondays in the year. The problem cited is typical of temporal series problems, and *the principle that should guide one in pooling results should be to group as wide a range of data as are typical with respect to the characteristic under investigation, but not affected by other seasonal or systematic tendencies.*

(b) Let it be required to ascertain the nature of the seasonal fluctuations of the load. In this case a tabulation by weekly units would be the best, as this would completely suppress both Saturday and Sunday and hourly idiosyncrasies. With this in mind, it is seen that a tabulation by six or eight-day or monthly periods would not be as satisfactory as weekly or bi-weekly periods.

The principle to follow is to use such a temporal unit as equals or is an integral multiple of the period within which occur the tendencies which it is desired to suppress.

Without claiming that ratios, periods, trends, etc., do not occur, or are unimportant, in other than temporal series, still it should be noted how peculiarly intimate is the statistic of importance and the series type. This intimacy is apparent whether graphic or algebraic treatment is involved.

The geographic series is much less amenable to algebraic treatment than the temporal, because the issue is connected with a place, not with a date which may be represented by a one-dimensional time continuum. An equally simple and well-known algebraic representation for geographical position is not available. The tightness with which the issue is connected with geographic position is the crux of the geographic series. A lessening of the bond should never be attempted. Accordingly, the two-dimensional map becomes the basic instrument in the treatment of geographical series data. The richness and accuracy with which data may be presented upon a map is nicely shown by the maps of the Coast and Geodetic Survey. No parallel neatness in algebraic treatment is available, but one statistic, the center of population, based upon a two-dimensional distribution of cases, has been worked out algebraically, thus becoming a statistic which is unique to the study of geographical series.

The typical qualitative series is obtained when alternative characters appear as a result of a single sort of sampling. Thus, if we ascertain and record the nationality of each person passing through a turnstile to see a baseball game, a number of nationalities,—qualities, categories,—would be recorded with varying frequencies in each. These qualities are not es-

sentially connected with moments of time or positions in space, for, though noted at a specific time and place, they maintain after the subjects have gone to a six o'clock dinner.

Furthermore, there is no obvious quantitative relationship between nationalities. What has one learned by such a tally? The numbers in each category (a term used only when rubrics are discrete) or class (a term used when rubrics are either discrete or continuous). The numbers or the relative numbers in each class and the relationship of these from class to class are the only issues involved. Such a concept as the mean nationality does not enter in. What is the mean nationality of three Germans, four Irishmen, two Frenchmen, one Japanese, and seven Americans. *The essence of qualitative data is that the items in it are not differences in amount on a single continuum.*

To the biologist the concept of gene or the allelomorph is essentially that of the qualitative series antecedent to the mapping of the chromosome, after which these become concepts of a one-dimensional spatial series. This nicely illustrates a common characteristic of qualitative series. The items frequently are qualitative because of inadequacy of one's knowledge as to spatial or quantitative relationships between classes. *One of the most fruitful lines of attack of qualitative data is the attempt to discover a sense in which it is not qualitative.* However, if no such discovery is made, there still remains the possibility of study of the series through the statistics essential to this type, namely, frequencies in a class, proportions in a class, and ratios of such proportions.

Lastly we have quantitative series, always presenting two dimensions capable of quantitative study: *(a)* different amounts of some single thing, and *(b)* frequencies of occurrence of these

different amounts. In the study of such series we obviously have the possibility of all the issues connected with qualitative series, for, neglecting the quantitative relationship of the variable in question, we have data recorded in classes so that *all qualitative series techniques apply without any modification whatsoever.* The beauty of the quantitative series lies in that a much greater variety of problems presents itself, a much richer, more exacting, and a more descriptive set of statistics are available for the analysis of relationships inherent in the data. A typical quantitative series would result if the scores of a certain class on a certain achievement test are gotten. A record of each score, together with the number of pupils receiving it, constitutes the basic series.

Life's problems do not confine themselves to single series, and certain methods have been developed for handling problems which are complexes of two or more of the four types mentioned, but it is well to recognize that in general the problem and the method are functions of a single series.

SECTION 3. TYPES OF STATISTICAL PROCESSES

The various processes of statistics may be listed in the order in which they take place: *(a)* definition and analysis of purpose, *(b)* collection of raw data, *(c)* tabulation of raw data, *(d)* computational procedures, *(e)* presentation of results through the employment of summarizing statistics or graphic devices.

(a) **Definition and analysis of purpose**: Antecedent to the initial collection of facts is the total life experience of the investigator in the light of which he has defined his problem with such detail that the hypothesis upon which he is building will be supported if the data yield

certain statistics,—or, more rarely, but generally with greater scientific insight, he has multiple hypotheses, each in turn associated with different or alternative statistical outcomes.

(b) The collection of data: The greatest care must be exercised to insure that the sample dealt with is "fair." It, of course, should be chosen because of some principle of sampling in mind, and the care should be that this, and not some unthought-of basis, is actually operating. This question may always be reduced to that of the comparability of the sample chosen, as observed, measured, or tested, with samples to be selected and measured in the future, or with samples that have been selected and measured in the past. The word "fair" implies a comparison. Thus, not one sample but at least two, and generally more, should be in mind,—on the one hand the immediate sample which the experimenter is about to use, and on the other hand the sample or samples with which comparison is to be made. Between a sample and its counter-sample, or contrast-sample, or over-and-against-sample, all conditions should be as nearly identical as possible except the one whose influence is being sought, that is, except the one under investigation. The existence of this counter-sample is inherent in all problems, though not specifically stated. Suppose one's expressed purpose is to find the mean 16-year-old level of accomplishment upon a certain psychological test. Assuredly the value in doing this would be because it enables comparison with other groups or individuals. If this counter-group is of 15-year-olds, then they should differ from the first group in that respect only, and not in such sundry other respects as sex, nationality, etc. The burden of insuring similarity between these two groups should not be entirely or mainly upon the shoulders of the person collecting the second group. It is essential at the time of the col-

lection of the first sample that the nature of subsequent contrasting samples be in mind, and the more definitely the better. It is incumbent upon the person collecting the first sample to define it completely and fully. Failure to do so is irremediable, for no care exercised by the collector of a subsequent sample can allow for undefined conditions in the first sample.

A different type of ability is called for to insure the proper collection of data from that necessary to its subsequent evaluation. For example, in securing scores upon a psychological or achievement test the collector should be sensitive to individual differences of his subjects which are abnormal, or more accurately, other than those assumed to hold. He should be sensitive to any unusual conditions under which the test is being administered, such as conditions of light or state of discipline, and he should be sensitive to temporary attitudes or conditions of the subjects which are unusual. None of these types of sensitivity are called into play in the computational work which follows. It should not be assumed that because a person is a good statistical computer he is also a good collector of data.

(c) The tabulation of data is usually a tiresome and routine undertaking. There is, in the first place, the assembling of the collected quantitative and qualitative items. Though the items have certain definitive individual characteristics, such, for example, as the name of the subject measured, the specific time and place gotten, these are usually considered irrevelant to the issue, and are not retained in tabulation. Just how much should be retained and how much discarded should not be decided upon the basis of the immediate needs of the investigator, but rather in the light of presumable or probable interests and needs of future investigators.

Investigators generally err by publishing too little original data. Examples of such failures are frequently present when sex of subject, or nationality, or even age, is not tabulated and made available to the reader, though each of these items is a part of the original record.

There should be an initial tabulation of raw data which includes, in addition to the items which will be used by the investigator, such other items as are available and as may be pertinent to the problem of workers in related fields. Then follow intermediate tabulations, choosing and organizing data for the purposes of the investigator. And, finally, there are tabulations of the findings,—terminal statistics,—of the study. Questions connected with these various tabulations are treated at greater length in Chapter III.

(d) **The statistical processes involved in the evaluation of data** range from such simple operations as making frequency tables and plotting distributions, time curves, etc., to the involved algebraic undertaking of deriving appropriate formulas and standard errors for the particular statistics being dealt with. Most of this text is concerned with the issues connected with the algebraic treatment and evaluation of data.

(e) **The presentation of results**: No matter whether the magnitude shown is represented by a distance (graphic portrayal), a numerical amount (quantitative series), or represents a frequency in a class (qualitative series), *there is always (1) some stated or implied counter-magnitude to which it is compared, and (2) some degree of credibility of the difference between the magnitude and its counter-magnitude is reached.* The most precise way in which this credibility is stated is in terms of the probability that a difference as great as that observed could have arisen as a matter of chance. Concepts *(1)* and

(2) are implicit, when not explicit, in every statistical evaluation. For example, to say that a child's mark in reading is 74 becomes meaningful only when some such added concepts are present as, first, the child is a fifth-grade pupil in a school wherein the passing mark is 70. There is also needed some idea of the variability of fifth-grade pupils in their ability; and, finally, the idea that the 74 is not a perfect representation of the ability of the child in question. To precise concepts of the sort mentioned should adhere a further precise concept,—that giving the probability that 74 is sufficiently above 70 that chance does not account for the excess. The method of reaching a precise statement of probability cannot be given at this point, but it should be obvious that some such concept, precise or indefinite, is inherent in the interpretation of the trustworthiness of, and therefore the significance of, every statistical difference. The reader should develop the habit of appreciating that *an answer in terms of probability exists for every observed difference between two statistical measures*, even though he does not have the necessary ability, or perhaps data, to arrive at a precise statement of it. Just as the error in statistical statements is ubiquitous, so should be the awareness by the student of this fact.

If the differences are great, as, for example, would be the case if the height of a normal adult man is compared with that of a small child, and if the fallibility of the measures is small, as would be the case if the height of each is gotten by precise methods, then the chance that the man is not truly taller than the child but merely seems so, due to the inaccuracy of observation, is of the order, perhaps, of one in a million. The first concept is that of the difference, and the second that of its trust-worthiness. Here

the second assumes a minor role, but the reader should note that even here it is present. Further, it is important to note that differences of this obvious sort are such as society has long recognized and become adjusted to, and they thus possess no particular interest. The interesting issues quite invariably attach to differences of such amount that a certain element of uncertainty as to trustworthiness persists. The crude techniques of the remote past have established such facts as that adults are taller than children, but it has remained for refined procedures, carefully expressing differences in terms of the chance that they may have arisen as a matter of chance, to indicate that, for example, the visual imagery of children is more variant than that of adults. The former issue,—adults taller than children,—is so well established that it has ceased to be a problem, while the latter is still a matter of interest, and thinking about it involves both the concept of a difference and the certainty or trustworthiness with which it is established. *These joint concepts are the two terminal points of statistical procedure.*

SECTION 4. THE SERVICE RENDERED BY ELEMENTARY AND BY ADVANCED STATISTICS

The evaluation of data may involve the range of statistical processes from the most elementary to the most abstruse. There is on the one hand a resort to statistics to make specific and quantitative a concept such as an average, and on the other hand to elucidate a relationship which the investigator no more recognizes as statistical than he does concepts of color, musical harmony, or beauty of form. On the other hand, concepts may be developed *de novo* from the data at hand and from a consideration of mathematical relationships which were not in the mind

of the investigator at the initial stage of his study. The making specific and the testing out of these concepts represent a functioning of statistics in its more powerful aspects. The elementary student may well devote his effort to finding appropriate mathematical or statistical expressions for concepts or issues which have been suggested to him from nonstatistical sources. This calls for ingenuity and leads to a development in thinking. As greater mastery and confidence in adapting statistical techniques to questions in mind develop, the student may find a reverse process occasionally takes place, and that issues are suggested by mathematical and statistical relationships discovered, and that these issues demand not only careful logical analysis, but quite frequently the collection and evaluation of new data. When this is the case, statistics is no longer a tool but also a source of inspiration. Derivations of fundamental formulas are the essential elements in training in this latter stage, but they will play a minor role in a first course. It is to be hoped that the reader will be annoyed by the paucity of derivations in this text, for that is a sign that the subject is assuming a place in his processes of mental analysis and is not merely a set of mechanical rules of operation.

The purpose of classification of series into types may not be apparent to the student at this stage of his study. As the study of graphic portrayal, averages, measures of variability, frequencies and proportions in classes, etc., proceeds, it will be found that *there is a chain of dependent techniques and of terminal statistics, graphic or arithmetic, which are consequent to the type of series dealt with.* Thus, as soon as the series is classified as of a type, the issues which are presumably of importance and the techni-

ques for working them up and presenting them are suggested.

It is true that much data may be classified in more than one way. For example, the achievement scores of pupils of different ages in different schools of a city may be classified as a quantitative series, a geographic series, or a temporal series. If the first, one variable is score and the other is frequency (age and school being neglected); if the second, one variable is school location as given on a map, and the other, achievement score—or the other might be number of pupils of a given age range or score range; if the third, the assumption is made that those of a given age are the same sort of individuals so far as achievement is concerned as those one year younger will become a year hence. Due to this assumption, this should be called a pseudo-temporal series, but the issues of interest, mental growth, etc., will be the same as those in a true temporal series. These three ways of looking at this simple body of data do not exhaust the points of view that may be taken. The student should realize that *the series type into which the data fall is nearly as much a question of point of view from which viewed as it is a question of the intrinsic structure of the data.* When data may be viewed as of more than one type, it frequently occurs that one or more of these ways of looking at it are more or less trivial, raising and leading to the analysis of no important questions.

A more detailed classification than into one of four is useful. The following has been made because in each instance one or more unique characteristics of treatment, or of evaluation, are connected with the type in question.

Temporal

T-t = temporal series in which increases and/or decreases, not necessarily

monotonic, are present. Include here T-t-r and T-t-i, temporal series involving ratios and indexes.

T-c = temporal series in which periodic or cyclical phenomena are present.

T-g = temporal series in which growth (monotonic) is present.

Spatial

G-c = geographic series wherein there is continuous variation in the variable from one geographic location to a neighboring location.

G-d = geographic series in which there is discontinuous variation in the variable from one geographic location to a neighboring location.

S = spatial series or one based upon location in three-dimensional space.

Qualitative

Ql-f = frequencies in qualitatively different classes.

Ql-m = magnitude of some function for qualitatively different classes.

Quantitative

Qn-c = frequencies in quantitatively different classes of a continuum.

Qn-d = frequencies in quantitatively different classes of a discontinuum.

Qn-o = frequencies in the ordered classes of a continuum.

PROBLEMS

As a vocabulary review the student may test his knowledge of the following words and phrases:

basal year
categorical series
center of population
character
characteristic
collection of data

continuous
continuum
counter sample
contrast sample
cycle
discrete
fair sampling
frequencies in a class
geographical series
growth series
index
natural limits
periodicity
price index
price ratio
price relative
probability
pseudo-time series
qualitative series
quantitative series
ratio
sampling
series
significance
spatial series
statistical series
tabulation of data
$\left.\begin{array}{l}\text{temporal}\\\text{time}\end{array}\right\}$ series
terminal statistics
trait
trend
zero point

Think of series of the sorts T-t, T-c, etc., of preceding Section 4. In one or two sentences describe each. The effort should be made to think of simple situations. When giving an illustration of one sort attempt to choose data which are so simple that no other series type is important in connection with it.

CHAPTER III

STATISTICAL TABLES

SECTION 1. CHARACTERISTICS OF A GOOD TABLE IN GENERAL

The chapter which follows deals with graphic methods and is concerned with charts, diagrams, graphs, etc., constituting pictorial representations of statistical series. The statistical table is quite different. Its purpose is not directly to give a picture of a sequence, but to provide the basic data from which such a picture, or at least the outstanding features of such a picture, may be determined and visualized if desired. The statistical table is simply a shorthand statement of facts. If a thousand or so facts of the sort, "The population of Aaber County is 4000;" "The population of Anthony County is 3200;" "The population of Avery County is 4800;" etc., etc., are to be presented, they can not only be more concisely shown by tabulation, but several thousand additional facts, such as "The population of Anthony County is 800 larger than that of Aaber County" are presented at the same time and in an agreeably compact manner. The desire to accomplish double, triple, or manifold

presentation by a single tabular arrangement is the desideratum which imposes conditions and determines appropriateness of procedure.

The same facts in regard to population are shown in the following five tables, and while not exhausting the possibilities of presentation, these will suffice to show the wide option which exists in presenting very simple data.

TABLE III A

Populations and Areas of Counties

Counties	Population 1920	Area in Sq. Miles
Aaber	4,000	480
Anthony	3,200	400
Avery	4,800	800
Bascomb	16,000	700
Brown	3,000	600

TABLE III B

Populations and Areas of Counties

Counties	Area in Sq. Miles	Population 1920
Aaber	480	4,000
Anthony	400	3,200
Avery	800	4,800
Bascomb	700	16,000
Brown	600	3,000

TABLE III C

Counties arranged according to Population

Counties	Population 1920
Brown	3,000
Anthony	3,200
Aaber	4,000
Avery	4,800
Bascomb	16,000

TABLE III D

Counties arranged according to Population

Counties	Population 1920
Bascomb	16,000
Avery	4,800
Aaber	4,000
Anthony	3,200
Brown	3,000

TABLE III E

Counties arranged according to Population

Population 1920	Counties
16,000	Bascomb
4,800	Avery
4,000	Aaber
3,200	Anthony
3,000	Brown

As judged by a single purpose, no two of the tables given are equally meritorious. If the table is to be used more frequently in abstracting information about various counties than as a means of comparing counties, i.e., if it is a reference table and not one pointing some conclusion, the items in the stub (the first column) should be arranged alphabetically, as in Tables III A and III B in order to facilitate the finding of items desired. If populations are more likely to be studied than areas, Table III A is preferable to Table III B, as the Population column holds a dominant position in Table III A.

Should it be intended that the table be not primarily a reference table arranged to simplify the extraction of items of information, but, let us say, to point conclusions with reference to populations, Tables III C, III D, or III E are preferable to Tables III A or III B. If counties of large population are the chief consideration, Table III D is preferable to Table III C, as the first row of a table ranks higher in dominance than successive rows. Next in importance is the last row. Totals or averages are, because of their importance, frequently placed in the first row, but if other items demand this position or if captions (headings of columns) are less readily interpreted when separated from the body of the table by a row of totals or averages, then the bottom row may be used.

As a means of pointing conclusions dependent upon populations, Table III E is to be preferred to Tables III C or III D, as the population data hold the dominant position in Table III E.

In general one should so draw up the table that the items in the stub and the captions constitute the argument or information with which the table is entered, and so that the column and row next to the stub and captions contain the most important items to be obtained from the

table. Rows and columns more removed from these dominant positions should contain less important data, except that the last row and last column may be given to data of first or second importance.

Such Tables as III A and III B are *primary or general-purpose tables*, since they *contain the raw data without abridgment, and may be used for various purposes*. Such *Tables* as III C, III D, and III E are *derived from primary tables*, such as III A and III B, and by *emphasizing certain facts serve a special purpose*. These two types of tables should be recognized. The special-purpose table is always published because it conveys the point of the study. The general-purpose table should always be published also, as it provides the only means of checking the author and of discovering if other or further conclusions can be drawn. Several tables and many calculations may be involved between the primary and the final derived table. If full description of these intermediate steps be given it is not essential that these intervening tables and calculations be published.

We have indicated that *there are three types of tables, (a) the general-purpose table*, or table presenting original data without intent to point a conclusion or emphasize some feature at the expense of others; *(b) the intermediate table*, or table which is derived from the general-purpose table in the process of putting the data in such form as to lead to a conclusion; and *(c) the special-purpose table*, or table derived from *(a)* and *(b)* so as to bring into relief some conclusion. Now let us briefly note what constitute desirable characteristics of all three, and then of each separately.

Principles of dominance and labeling are the same for all three sorts of tables. The approved lettering is such as can be read if the observing

eye is in front of the bottom of the page. If demands of typography require vertical lettering, it should always be so as to be read from the right-hand margin, that is, the eye is considered to be at the right, and not at the left. This principle of location of the observing eye at the bottom or right applies equally to lettering, whether horizontal, vertical, or oblique, of graphs. The title should be at the top of the page. When placed at the bottom, as is occasionally done, it is in a position of second dominance. *The title should be brief and answer the following question*, which we may always assume is in the reader's mind, "*Is the subject matter of this table such that I am interested in looking at the table?*" If brevity of title is incompatible with accuracy, there may frequently be employed a subtitle in smaller type, or a footnote, thus preserving the brevity and large type-size of the main title. It is the thing that the eye is supposed to rest on first as the page is viewed. This same principle applies to titles of graphs.

There are always two means of entry into the table. One is given by the items in the stub (the left-margin column of the table) and the other by the items in the captions of the columns. Generally the means of entry involving the larger number of items is made the stub, so that the smaller number of items may be indicated in the captions. The column next to the stub is the dominant column, while the next in dominance is the last column or one at the extreme right of the table. The dominance of any column may be increased by enclosing it in heavy rules. The row next to the captions is the dominant row, while the next in dominance is the last row. The dominance of any row may be increased by heavy ruling or leading.

SECTION 2. SPECIAL CHARACTERISTICS OF THE THREE TYPES OF TABLES

General-purpose table: There are certain special requirements of the general-purpose table. It is to present the original data in all its detail except for items considered to be immaterial. For example, if John Doe, age 13.5, seated at desk 17 of Mary Brown's fifth-grade class, secures a test score of 86, it may well be that the general-purpose table merely records the sex, the age, the grade, and the score, it being assumed that the name of the boy, that of his teacher, and the desk where seated are irrelevant with reference to any of the sundry purposes with which readers may approach the table. The author may be incorrect, for some of his readers might be interested in, say, the teacher as related to pupils' scores. However, this may be, it is clear that it is incumbent upon the author to anticipate such potential uses of his data, and in the light of the variety of usage that he can anticipate, construct his general-purpose table. He should seldom limit the presentation so as to cover the single type of use to which he himself has put the data. If he does so limit it, then the table is still of value in permitting a reader to verify the author's computations, but this is probably a much narrower value than could have been given to it.

We may not say that in publishing a general-purpose table the author should have no purpose, —rather, he should have a variety of purposes, as many as the uses that he can anticipate or imagine. He should subordinate his own special purpose and serve these many purposes in the arrangement of the table.

Of first importance is consideration of the means of entry which his reader will be called upon to use.

The devisor of the general-purpose table must think of the concepts and classifications which he can count upon finding in the minds of his readers. He can certainly count upon their knowledge of the alphabet. He can count upon their knowledge of chronological order, and, with somewhat less certainty, knowledge of certain of the major geographical relationships. He can believe that certain classes such as "live stock," "cereals," "physical measurements," "mental measurements," "maintenance costs," "instructional costs," etc., will be intelligible and will include about the same items for all of his readers. In general, the arrangement and classifications of stub and captions are to be such as call upon knowledge of the sort mentioned. Alphabetic entry is, in particular, of wide utility. Such emphasis as is consequent to the arrangement of the table should be for the sake of ease of entry to the table, and not to emphasize some item therein.

A logical arrangement involving the concepts of subordination and coordination is usually a psychological aid to entry. For example, captions such as these

FARM PRODUCTS

LIVE STOCK					FRUIT			
CATTLE	HOGS	SHEEP	HORSES	OTHER	APPLES	PEACHES	PEARS	OTHER

provide a number of co-ordinates under a single superordinate. The ruling and type size bring into relief the classification used and required of the reader in order properly to enter the table. Obviously the principle to follow is to *make rows or columns contiguous that will generally be used together.* Another principle to be kept in mind, though it frequently must be vio-

lated, is that normal reading habits and muscular development of the eye favor horizontal eye movement rather than vertical.

In the general-purpose table there should be emphasis on the natural arguments used by and the presumed interests of the reader, while *in the special-purpose table there may well be emphasis on the argument used by the author and upon his special interest.* In the general-purpose table rulings, leadings, and placing in dominant positions should be in the light of the reader's already established modes of thought, thus nothing new is being taught him, while in the special-purpose table these same things are used with the view to teaching him something, or giving an emphasis that is new to him.

Special-purpose table: The essential characteristics of the special-purpose table have already been indicated. It is designed to point a conclusion. It generally consists of derived statistics and presents a difference between two or more statistics. It may well be that there are many intermediate processes of assembly and computation between the items of the general-purpose table and the resulting special-purpose table. A sample of these intermediate processes, or a very thorough description of them may suffice. In any case, sufficient should be given so that the doubting reader will be able to start with the data of the general-purpose table and verify all the statistics given in the special-purpose table. It is obvious that the special-purpose table is always published, for it constitutes the conclusion made by the investigator. There is equal requirement in an article claiming scientific value that the general-purpose table be published. It may sometimes suffice in semi-scientific publications intended for popular consumption to forego publication of the original data, if reference is made to some source, perhaps

a more technical journal, where it is available. This procedure is a makeshift and not in general to be recommended.

In the desire to place emphasis, great care must be exercised to prevent distortion. For example, a table admittedly emphasizing the size and importance of Myberg might be as follows:

Cities	*Population*
Myberg	**2174**
Ladome	14134
Momcup	11655
Defton	6758
Frilty	2084

The emphasis upon Myberg has been accomplished by recording in the dominant row. This, in itself, is not to be critized, but emphasis by recording the population in more expansive type horizontally than that used for the cities with which Myberg is being compared is misleading, or if the reader discounts the type size so accurately as not to be misled, it nevertheless has been annoying to him. A better presentation is as follows:

Cities	*Population*
Ladome	14,134
Momcup	11,655
Defton	6,758
MYBERG	2,174
Frilty	2,084

Here again the author has emphasized Myberg, but he has now done so without sacrificing accuracy. The significance of the population of Myberg is only sensed when the contrast statistics are known. Thus it is equally important that the reader know the population of the other

cities. These should accordingly be given in such type and in such arrangement as best to enable comparison with the population of Myberg. Whereas the item of entry to a general-purpose table has been left to the particular interest of the reader, an endeavor has been made in the special-purpose table to dictate this, in so far as the author is able to do so, by utilizing the available means of emphasis. Though the author may legitimately dictate the issue to be attended to in the special-purpose table, he is never justified in misrepresenting any quantitative relationships inherent in the data, or even in suppressing by omission or relegating to obscure positions data which are of most comparative importance.

Intermediate tables: Such intermediate tables and computations as have been employed in reaching the conclusions shown in the special-purpose table, or in other terminal statistics, represent the development of the plot, as it were. The author of a short story may revise and refine some elementary plot a score of times before it finds expression in his published work. The more hidden, though sound, the steps in its development, and the cleverer it is, the more appealing to the reader. Unhappily, in the scientific study no concealment of the steps of development is permissible. These should be fully described and, if quite involved, illustrated by examples which may include both tables and computations. The scientist is called upon to belittle his own genius by a full demonstration of the directness and simplicity of his processes. If he does otherwise he may be fawned upon by the Ladies' Aid or the Up-and-Coming Dinner Club, but he will hardly aid the efforts of serious students. The student should remember that verifiable knowledge is the object of his endeavor, and the verification should be by others. Common experience

indicates that occasionally there are capable students whose enthusiasm or whose distaste for laborious presentation leads them to jump from finding to finding without stopping to prove their course, with the net outcome of detachment from, and lack of influence upon, fellow students in the same field.

It happens not infrequently that some statistic readily derived from the data given in a general-purpose table will be used so generally as to warrant its inclusion in such a table. For example, if the population of the states of the United States constitutes the data of an original table, there surely should be recorded in this table the population by divisions, and finally the population for the United States entire. These are derived statistics and do not constitute basic items in a general-purpose table. It may be believed, however, that they will be used almost universally by the readers of such a table. It is thus a courtesy to the reader to make the necessary additions for him, and to record these totals. In so far as this is done, the table has taken on a special-purpose function,—one amply justified by the generality with which these totals will be desired by readers in pursuing their separate interests. When such derived statistics are of such importance as to warrant inclusion in what is otherwise a general-purpose table, they generally warrant being placed in a dominant position, the row next to the captions, or the column next to the stub, or the last row or last column. They may even warrant inclusion in a bolder type-face or with different leading than the more detailed or primitive data which are basic to the table. Table IV S illustrates this, and it also gives the standard of U. S. Census practice in dividing the United States into divisions and in the order in which the divisions and states are listed. The most common

derivatives which may add to the usefulness of a general-purpose table are totals, arithmetic means (or other averages), standard deviations (or other measures of variability), proportions, and percentages. The main principles involved in the construction of statistical tables here discussed, as well as certain other points, were found in Bowley (1926), Crum and Patton (1925), Day (1920), and Young (1925).

PROBLEMS

Problem 1. The student may test his knowledge of the subject matter of this chapter by giving himself a vocabulary test involving the following words and phrases.

general-purpose table
special-purpose table
intermediate table
original table
derived table
dominance
order of dominance of portions of tables
stub
caption
observing eye
terminal statistics

Problem 2. Draw up a general-purpose table, presenting data as here described: In connection with a certain study the following items of information for 300 American white delinquent boys in a Texas juvenile training school, ages 8 to 20, were gathered. The order in which they are here mentioned is haphazard.

name
height in mm.
length of head in mm.
strength of grip of right hand in lbs.
age in completed years and days of a partial year
score on a constructive ability test
speed of tapping (30 seconds with the preferred hand)

strength of vision of left eye - Snellen chart test
circumference of head in mm.
lung capacity in cu. inches
school grade
handedness (left or right)
strength of vision of right eye
strength of grip of left hand
number of scars on cranium
mental test score
breadth of head in mm.
pubertal development (pre-pubescent, pubescent, or post-pubescent)
weight in pounds

Assume that the users of the table do not know the names of the boys, so that an alphabetical arrangement by name of boy will not be serviceable. The steps to follow are:

1. Choice of title with or without subtitle.
2. Selection of the two main lines of classification, and allocation of one to the stub and the other to the captions.
3. The classification of items within the stub in harmony with the knowledge of the reader.
4. A similar classification of items in the captions.
5. An arrangement of items of stub in conformity with the principle of placing the most generally used items in dominant positions, and in keeping related items together.
6. A similar arrangement of the items of the captions.
7. The use of special rulings, leadings, type size, bold face, italics, and capitalizations to facilitate entry as it will presumably be made by the reader.
8. The inclusion or non-inclusion of certain derived measures in the light of the gener-

ality with which it is expected they will be used.

9. The actual lettering of title, stub, and captions in order to supplement ease of entering, as planned for in the preceding eight steps, and in harmony with the principle of the location of the observing eye.

Problem 3. Draw up a special-purpose table, using the accompanying data taken from Brigham, Carl C., *A Study of American Intelligence* (1923), p. 120. Your purpose is to show comparative intelligence test scores for certain Nordic and Mediterranean strains, having defined country of origin as representing strains as indicated below. Mean army intelligence test scores of certain drafted men according to country of origin were found to be as follows:

Nordic: Belgium 12.8, Canada 13.7, Denmark 13.7, England 14.9, Holland 14.3, Norway 13.0, Scotland 14.3, Sweden 13.3.

Mixed: Austria 12.3, Germany 13.9, Ireland 12.3, Poland 10.7, Russia 11.3

Mediterranean: Greece 11.9, Italy 11.0, Turkey 12.0.

What statistic or information in addition to that given in the preceding problem do you consider would be of prime value in serving the stated special purpose? (The answer to this question may appropriately be postponed until the next six chapters have been studied.)

CHAPTER IV

GRAPHIC METHODS

SECTION 1. GENERAL FIELDS IN WHICH SERVICEABLE

An essential difference between a written description of a landscape and a painting of it is that the reader exercises minimum initiative in determining what is noted, while the viewer of the painting, no matter how subtly and successfully the high lights have been drawn, does take an active and personal part in determining the focus of attention. The wider the experience and training of the observer the more likely that the feature of most worth in the painting will be apprehended. For the fullest enjoyment not only must the artist have been an expert in design and execution, but the observer must be expert also.

For fullest utility, the graphic portrayal of statistical data is dependent upon skill in presentation and commensurate skill in interpretation. When both of these are present the enjoyment and fullness of meaning gotten in reading a graph exceeds that in reading a description of the same data. We can, very reasonably, consider the meaning gotten from reading a

graph exceeds that in reading a description of the same data. We can, very reasonably, consider the meaning gotten from reading a graph as a product of d, m, and r, wherein d represents the data with its complete meaning, m, a quantity between 0 and 1, the excellence of the art of the maker of the graph, and r, also between 0 and 1, the expertness of the reader of it, thus $dmr<d$ if either m or r is short.

If the reader of the graph is quite untutored the maker must resort to stars, arrows, colors, heavy lines, light lines, shading, verbal addenda, pictograms, etc., to make the story clear and make the reader's task as light as possible. If the reader is inexpert he is correspondingly gullible and the maker is under a heavy obligation not to resort to the little tricks of bias, under and over-emphasis, so available and so misused by the promoter.

Various graphic devices are listed in Table IV A, arranged vertically according to felt ease of understanding, that is simplicity as the lay reader conceives it, and horizontally according to field served.

Distorted pictures, maps, and cartoons with a quantitative element may be designed to relieve the reader of necessity for search for and correct appraisal of the significant feature, but they may also be designed to suppress or exaggerate phenomena and to prevent correct appraisal. The demarcations between the inaccurate and the accurate, and between the popular and the technical, in graphic portrayal are hair lines which, though willfully crossed by promoters, are also often crossed inadvertently by amateur statisticians.

When the nature of the data permits, the picturing of facts conveys a readier comprehension than does a tabular array of figures. Since there are but two dimensions to the surface of a sheet

TABLE IV A
COMMON GRAPHIC DEVICES

	SERIES TO WHICH APPLICABLE			
	Quantitative	Qualitative	Temporal	Geographical and Spatial
Easiest to understand as judged (or mis-judged) by reader	Distorted pictures Pictogram Histogram Frequency polygon	Pictogram Circle chart Bar diagram Segmented bar diagram	Pictures showing changes with time, no precise labeling of either axis. Time chart	Map with pictures at sundry locations. Oversimplified map.
Of intermediate difficulty, requiring some explanation if presented to a popular audience	Ogive, or percentile curve	Very Simple 2-way plot. Block diagram.	Relative time chart. Composite price index or aggregate chart. Growth curve. Genealogical chart.	Stereoscopic presentations. Vegetation, rainfall, barometric maps, etc. Geodetic and Coast survey map. Road map. Architect's working drawing.
Understanding calls for a certain degree of technical training	Smoothed curve. Mathematically fitted curve. All presentations involving 2 or more variables in addition to frequencies: Correlation table, etc. Alignment chart.	Contingency table involving 2 or more vari-ables in ad-dition to frequencies.	Geological chart. Semi-logarithmic chart.	Mercator and other projections. Navigation and weather maps.

of paper, ordinarily but two variables are depicted in a single graph.

SECTION 2. TEMPORAL SERIES

Time chart: Consider the accompanying data giving the maximum temperatures recorded by the Weather Bureau for each day in July and August, 1917, for New York City.

TABLE IV B

MAXIMUM TEMPERATURE FOR EACH DAY IN DEGREES FAHRENHEIT

July 1-Aug. 31, 1917
New York City

July	1	80°	July	17	87	Aug.	1	98	Aug.	17	85
	2	88		18	80		2	96		18	80
	3	74		19	77		3	83		19	81
	4	78		20	83		4	80		20	84
	5	81		21	81		5	82		21	85
	6	80		22	86		6	82		22	80
	7	79		23	86		7	88		23	76
	8	70		24	86		8	78		24	83
	9	75		25	84		9	83		25	82
	10	65		26	85		10	80		26	74
	11	66		17	90		11	82		27	82
	12	71		28	80		12	83		28	80
	13	81		29	81		13	83		29	83
	14	81		30	95		14	78		30	81
	15	75		31	98		15	81		31	75
	16	85					16	80			

If it is desired to study diurnal changes in maximum daily temperatures, a graph is made in which the abscissa represents the days in order, July 1, July 2, etc., and the ordinate represents the temperatures, 0°, 1°, 2°, etc. For July 1 the ordinate is 80, for July 2, 88, etc. A line connecting the successive ordinates provides a

picture of the changes in maximum temperature throughout the two months.

CHART IV I

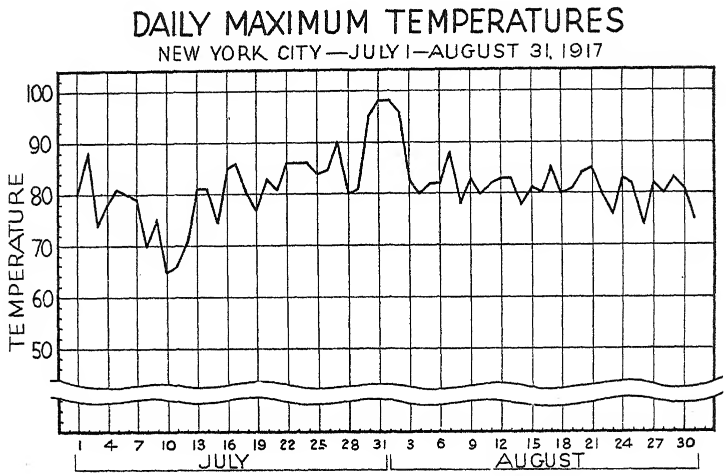


CHART IV II

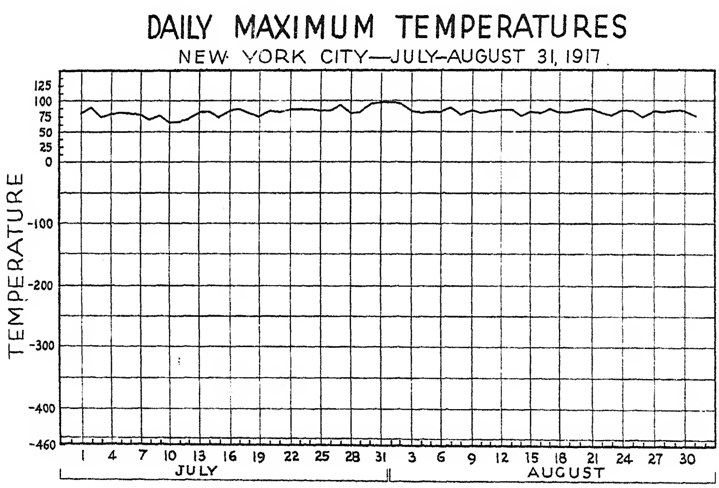


Chart IV I is such a graph and is called a Time Chart. Its two dimensions are time and magnitude of something as it changes with a change in time.

Certain detailed features of charts, zero points, scales used, rulings and proportions: As the initial and final dates do not represent the beginning and end of time, but are determined by the author's whim, *there should be no heavy vertical rules bounding the curve. A little space should be left between the initial plotted point and the left vertical rule of the chart and between the final point and the right rule of the chart,* as this gives the entirely correct impression that the curve is dangling in time and that both earlier and later observations were possible and might have been plotted. *The double wavy line near the bottom of the chart is drawn to convey the information that the* temperature *scale has been broken and that the actual zero point* on the temperature scale *is much lower down than indicated by the heavy horizontal rule at the bottom of the chart.*

The omission of a large portion of a vertical scale is sometimes dictated by limitations of space available. For accurate portrayal *it is desirable that a natural zero point be the zero point of the raw scores or measures and that graphic distances be correct representations of distances above this point.* Following this principle in the case of these temperature data, we measure temperatures from absolute zero, −460° Fahrenheit. The temperatures then appear: July 1, 540° above the zero point, July 2, 548° above, etc. The vertical dimension would then be some nine times as great as shown, unless this scale is reduced, and if reduced to 1/9 the Chart IV I scale we get Chart IV II, which is not very helpful to a prospective summer visitor to New York or to a salesman of air-conditioning equipment.

Natural zero points for one purpose may not be natural for another. For a physicist, $-460°$ may be a natural zero point on the temperature scale. For a banana planter, perhaps $32°$ is the important and, for his purposes, the natural point of reference. For a physician or hospital nurse, perhaps $98.6°$ is such a point. The writer will not attempt to define "natural zero point" for the definition must depend upon the particular phenomena and purpose in hand, but he would point out that there are such points and that they are the points from which scores or measures should deviate for the most meaningful presentation. In the field of prices an obvious zero point is \$0.00. *In situations where the top value is necessarily limited,* as for example is 100 per cent of a quantity, *this value becomes a natural point of reference.* There may be two natural points on the scale. In dealing with quantities that are percentages of a total, both 0 per cent and 100 per cent are such points. It is to be noted that in the matter of maximum temperatures $100°$ is not a limit and Charts IV I and IV II as drawn are not bounded at the top by a heavy horizontal rule. *Where there is such an upper limit a heavy rule representing it is desirable.* The ogive curve shown in Chart IV XIII is representative of this situation,—it being bounded at 0 and 100 per cent by heavy rules at left and right. If, for the artistic purposes of make-up it seems desirable to bound the curve with a rule at the top, it should be well above the maximum ordinate of the curve so as to give as little impression as possible of being a part of the data.

Ordinarily the vertical rulings of a time chart should be without emphasis, as one date is neither more nor less important than another. An exception can be made in the case of a relative time chart, as shown in Chart IV IV, where prices

or magnitudes are expressed relative to that at some basal date, the price or magnitude at this date customarily being called 100. *In this instance the vertical rule at this date may be drawn heavier than for other dates,* as shown in Chart IV IV. A second instance arises in connection with growth curves, for here the beginning, such as *the date of birth, is a natural starting point and may well be indicated on the chart by a heavy vertical rule.*

For all the common charts, except circle diagrams, the scales used are at the discretion of the maker. It has been claimed that the greatest artistic balance is reached if the ratio of over-all height to over-all width is the golden mean provided by the extreme and mean ratio. This ratio, $x/(1-x) = (1-x)/1$, holds when $(1-x)/1 = .618/1.00$, or approximately 3/5. *General experience supports the view that a pleasing result is obtained if the height of a curve is 3/5 its width.* In Chart IV II, if attention is fixed upon the curve only and not upon the bounding rules, the ratio of height to width is about one to thirty, which is altogether undesirable.

The lettering of a chart follows the same principles as those applying to a table. A chart may require an explanatory "legend" as shown in Chart IV IV. The upper left corner is an excellent place for such a legend, but it may be placed in any free region. *Vertical and horizontal scale rulings should be reduced to a minimum and given in very light lines,* except that some slight emphasis may be given to round-number rulings usually represented by every fifth or every tenth line.

Limitations of the time chart: Let us plot the several data of Table IV C upon a single time chart. The result is shown in Chart IV III, where three different ordinate scales have been employed, one for wages, one for steak prices,

TABLE IV C

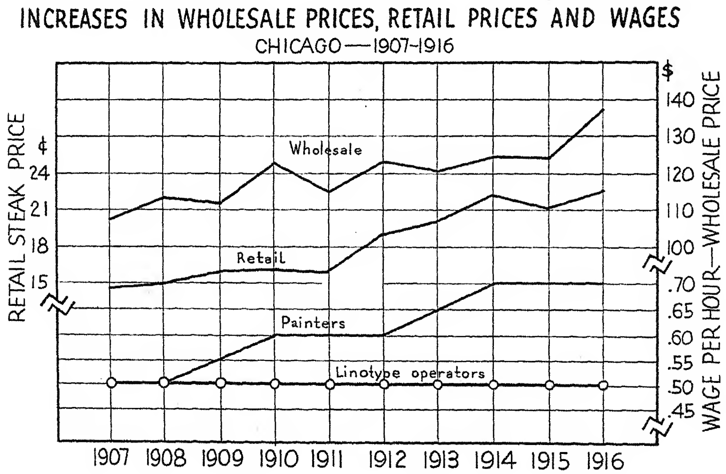
PRICE AND WAGE DATA

CHICAGO*

Year	U. S. Entire Dunn's Wholesale Price Index	Av. Yearly Retail Price Round Steak	Union Wage per Hour	
			Painters	Linotype Operators
1907	$ 107.264	14.3¢	50¢	50¢
1908	113.282	14.9	50	50
1909	111.848	15.9	55	50
1910	123.434	16.2	60	50
1911	115.102	15.9	60	50
1912	123.438	19.1	60	50
1913	120.832	20.2	65	50
1914	124.528	22.3	70	50
1915	124.168	21.2	70	50
1916	137.656	22.6	70	50

* U. S. Dept. of Labor, Bur. of Labor Statistics. Union Scale of Wages and Hours of Labor, 1916.

CHART IV III



and one for wholesale prices. The advantages of this presentation over one showing three charts lies in a saving of space and in nothing else. In fact, there is a decided disadvantage in having three such separate scales upon a single chart. Accurate comparisons between the wholesale, retail steak, and wage curves are not possible by this chart, though the existence of the three curves in a single chart encourages the reader to attempt to make them. The test of any graphic presentation is that the judgment of magnitude consequent to the visual impressions received is correct. By this test Chart IV III is nearly a complete failure,—the comparison of wage of linotype operators with that of painters being the only inter-curve comparison that is sound.

Not infrequently interest in, or, we may say, importance for our purposes of temporal data decreases as more and more remote time is considered. This decrease may be of much the same order as the decrease in size when things are viewed in perspective. When temporal data are shown we have a retrospective chart. It has points of similarity both with the semi-logarithmic chart and the block diagram, as discussed by Karsten and Brooks in *"Retro" Charts* (1943).

The relative time chart: The relative time chart is designed to enable comparisons of the relative changes in different magnitudes accompanying a change in time. It does permit of such comparison, but only with reference to some definite data which has been chosen as the basal date.

The basal year may be any date within or without the range of time presented. There are f ır common bases for the choice of one certain date rather than a second, and all of these have the merit of choosing a moment in time when the phenomena are, in some respect, more stable than at neighboring moments. Temporal phenomena have the habit of oscillating, though usually very

irregularly, as time changes. Such a function, if continuous, has a moment of no change at each maximum and minimum and a moment of uniform change at each point of inflexion. These points have features of stability which are advantageous in a point of reference,—the point of inflexion of a smoothed temporal series would seem to give an especially excellent and stable point for a study of trends. In economic phenomena the height of inflation, the bottom of depression, and a date between the two when the rate of change seems regular are common basal dates. The fourth reason for a certain choice is custom, which in turn is commonly related to one or another of the three conditions mentioned, though it may be so related in terms of much more varied data than that given by a single variable.

We will show increases of the prices and wages recorded in Table IV C relative to their values in 1907, which date becomes the basal year. We first compute, from the data of Table IV C the ratios as given in Table IV D and then plot as shown in Chart IV IV.

CHART IV IV

INCREASE IN PRICES AND WAGES—RELATIVE TO 1907

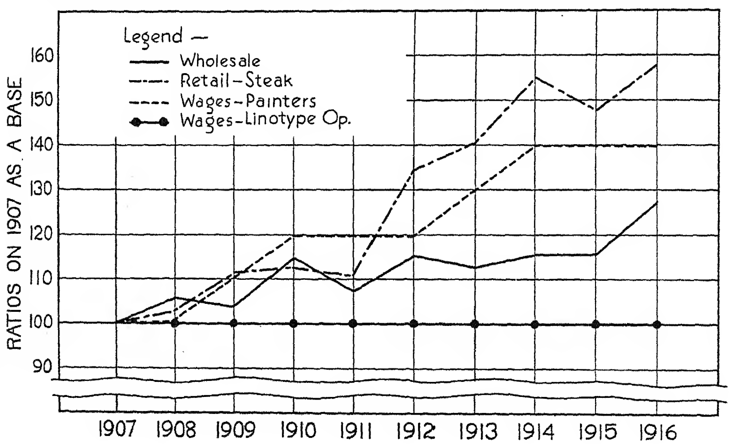


TABLE IV D

PRICE AND WAGES EXPRESSED AS RATIOS* 1907 AS BASE

CHICAGO DATA

Year	Dunn's Wholesale Price Index	Av. Yearly Retail Price Round Steak	Union Wage per Hour	
			Painters	Linotype Operators
1907	100	100	100	100
1908	106	104	100	100
1909	104	111	110	100
1910	115	113	120	100
1911	107	111	120	100
1912	115	134	120	100
1913	113	141	130	100
1914	116	156	140	100
1915	116	148	140	100
1916	128	158	140	100

* The decimal point is omitted, as is usual, so that a "ratio" of 106 means in fact 106/100.

The vertical rule for 1907 is made heavy because this date is different from any of the other dates, it being the basal year. Also the horizontal rule for 100 is heavy because this also is unique, all distances above or below being correct measures of percentage change from the 1907 status.

The relative time chart has certain inherent shortcomings. Changes relative to some other date than the basal date are frequently of interest, but are not shown, and second, a given distance below the 100 rule, representing as it does the same proportionate change as an equal distance above, becomes misleading. If the 1907 wholesale price, $107.26, doubles, it increases 100 per cent, becoming $214.52. If this price halves, it decreases by but 50 per cent. The

price must fall to \$0.00 to decrease 100 per cent. A change to \$214.52 and a change to \$0.00 are represented by the same distance on the relative time chart, but certainly the economic upheaval in the one case is in no sense equal, but opposite in nature, to that represented by the other.

Advantageous as the relative idea is, it is nevertheless grossly inadequate if great inequality in relative change of items is present. The relative time chart technique should be limited to situations wherein percentage changes are small, or at least of a somewhat uniform nature for the different variables involved. Where relative changes are large, a more informative graphic presentation, though one calling for greater knowledge upon the part of the reader, is by means of a semi-logarithmic time chart or semi-logarithmic relative time chart.

Semi-logarithmic charts: The characteristic of semi-logarithmic charts is that the scaling in one dimension, usually the ordinate, is proportional to the logarithms of a variable, while in the other dimension, which is frequently time, the variable itself is represented. If the scaling in both dimensions is logarithmic the chart becomes *a logarithmic chart*. Since, for all positive numbers, differences in the logarithms of numbers exactly parallel proportional differences in the numbers themselves, logarithms are peculiarly appropriate in studying situations wherein equal proportionate changes have equal psychological or economic significance. Semi-logarithmic cross-section paper is an aid in making these charts, as it is then not necessary to look up the logarithms of the magnitudes plotted. This aid is quite unnecessary as the looking up of logarithms prior to plotting is a very simple task. Even this small labor may be avoided at the cost of some accuracy if equal spacings upon ordinary cross-ruled paper are

labeled with numbers whose logarithms differ by approximately equal amounts as given in Table IV E for a thirteen- and a twenty-division scale. Another simple way to make one's own logarithmic paper is to scale blank paper according to the divisions on the slide of a slide rule.

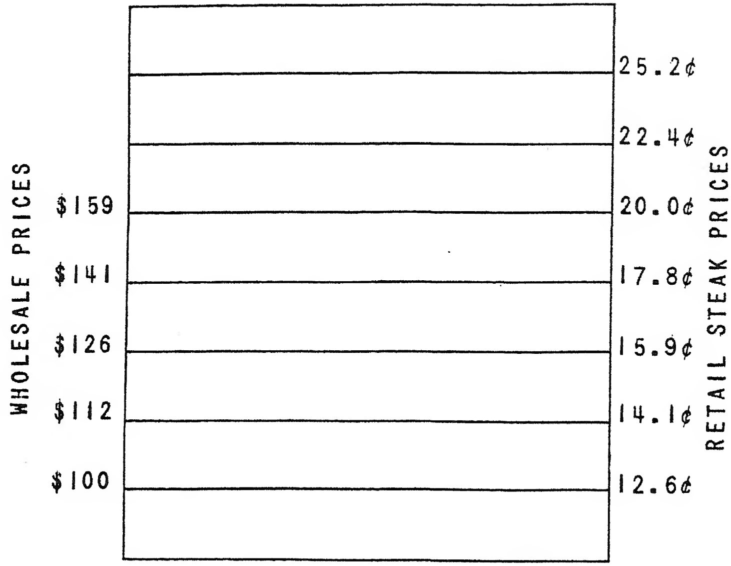
TABLE IV E

NUMBER VALUES CORRESPONDING TO EQUAL LOGARITHMIC DIFFERENCES

Thirteen-division scale to 2+ places		Twenty-division scale to 2+ places		Twenty-division scale to 3 places	
10	34.5	10	35.5	100	355
12	41	11	40	112	398
14	49	12.5	44.5	126	447
17	59	14	50	141	502
20	70	16	56	159	563
24	84	18	63	178	631
29	100	20	71	200	708
		22.5	79.5	224	795
		25	89	252	892
		28	100	282	1000
		31.5		316	

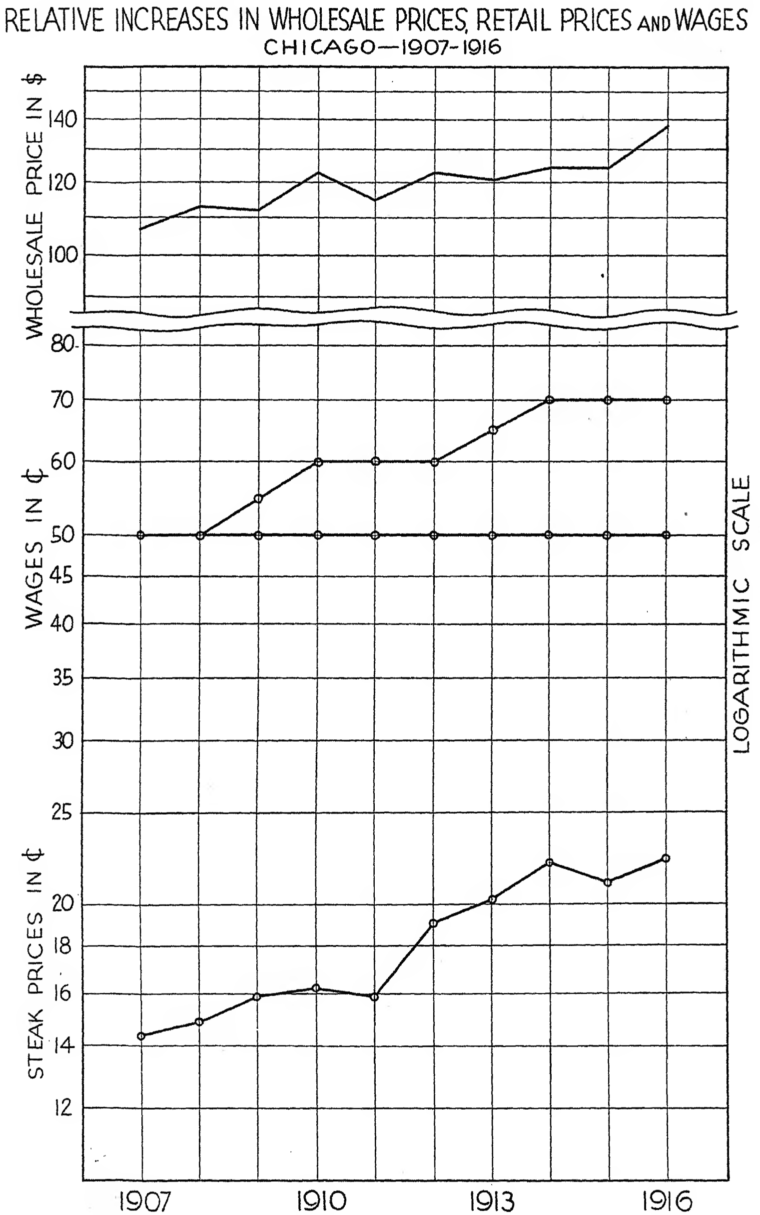
A benefit derivable from equally spaced rulings, labeled as in the thirteen- or twenty-division scales given is that two curves whose relative changes are important, but which would be widely separated if a single logarithmic scale is used. may be brought close together, using an equally-spaced grid and labeling differently for the two variables. For example, the wholesale price in dollars and the retail steak price in cents of Table IV C and Chart IV V can be shown by two curves in close proximity upon an equally-spaced grid by labeling a left-hand ordinate for wholesale prices and a right-hand ordinate for retail

steak prices, as follows:



The wage and price data are plotted on semi-logarithmic cross-section paper in Chart IV V. Any year may be chosen as a base and the changes in heights of the curves noted for any other year in which interested. The distances measuring these changes are correct representations of the magnitudes of the proportionate changes between the two dates. This flexibility in choice of the basal year and the accuracy with which proportional changes are shown are merits not possessed by the relative time chart. A disadvantage in Chart IV V as drawn is that a comparison of changes in different curves requires a comparison of linear distances that are considerably removed in space from each other. This difficulty can be avoided by plotting the curves upon tracing paper and then shifting each curve up or down as may be necessary until coincidence is attained for some desired date.

CHART IV V


RELATIVE INCREASES IN WHOLESALE PRICES, RETAIL PRICES AND WAGES
CHICAGO—1907-1916
WHOLESALE PRICE IN $
140
120
100
80
70
60
50
45
40
35
30
25
20
18
16
14
12
WAGES IN ¢
STEAK PRICES IN ¢
LOGARITHMIC SCALE
1907
1910
1913
1916

As a further illustration of the use of logarithms let us consider stocks having closing prices as follows:

TABLE IV F

CLOSING STOCK QUOTATIONS

	Previous day's closing price	This day's closing price	Change
Stock A	12 1/2	13	+ 1/2
Stock B	162 1/2	169	+ 6 1/2

Neither an examination of these figures not involving a computation with them, nor an examination of a time chart graphically showing the prices will be very helpful for such issues as ordinarily would interest one. The fact that Stock B increased $6 more than Stock A is probably not of importance. The proportionate change, not the absolute change, in price is the item of moment to economist or investor. The computation of the ratio of the price at one date to that at a second, and a comparison of such "price ratios" is a common practice and does serve an immediate need, but in general is quite inadequate in that the second date, or basal date, must be changed with each change in issue or interest. The issues become quite simple and interpretations very direct if one but thinks in terms of logarithms of prices instead of the prices themselves. For the quotations cited the data are:

TABLE IV G

LOGARITHMS OF CLOSING STOCK QUOTATIONS

	Logarithm of previous day's closing price	Logarithm of this day's closing price	Change
Stock A	1.0969	1.1139	.0170
	2.2109	2.2279	.0170

We immediately see that the proportionate increases are equal. Without any change in computation as basal dates in which interested change, any two prices may be compared and changes for different time intervals may be compared. The difference given, .0170, is the logarithm of the proportionate change in price. The number whose logarithm is .0170 is 1.040, informing us that the second price is 1.040 times the first, or that there is a 4 per cent increase in price in the case of both stocks. The relationships so clearly brought out by logarithms is graphically portrayed when amounts are plotted on coordinate paper with a logarithmic scale. The statistically-minded will hope that a daily paper will some day publish the logarithms of market quotations themselves. For many purposes, including writing checks and receiving wages, it is necessary to deal with the ordinary number system and such things as dollars and cents, but for the study of relative changes in things, such as prices, we are better served by the logarithms of magnitudes than by the magnitudes themselves.

Time Ratios: The quotient of a magnitude, such as a wage, a price, a measure of production, or a level of attainment, at one time divided by the similar measure at a different time is called a time ratio. A combination of such time ratios yields a general index, such as are the various business, cost of living, and price indexes. The properties of time ratios and of aggregates of them are fully treated by those making and using index numbers. It must here suffice to note that simple averages of such ratios generally have systematic errors of one sort or another due, among other things, to the generally skewed nature of distributions of ratios, as illustrated in Section III, Table IV N and Chart IV XII.

The Growth Curve: An important temporal series arising in biometric studies is one involving

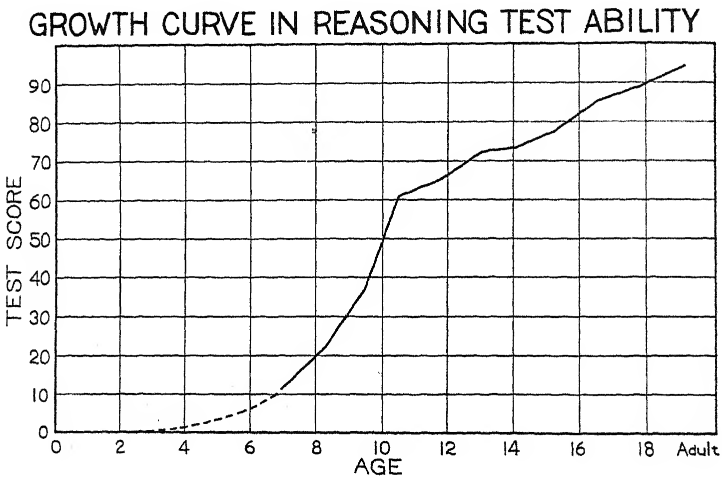
growth from the beginning, or near the beginning, to maturity.

The accompanying table gives smoothed scores in a reasoning test as given by Kelley (1917). Plotted they give a typical growth curve.

TABLE IV H

Age....	7.0	8.3	9.4	10.5	11.8	13.0	14.1	15.3	16.5	17.8	19.2
Score on Trabue Scale..	1.1	2.2	3.7	6.1	6.5	7.2	7.3	7.8	8.5	8.9	9.4

CHART IV VI

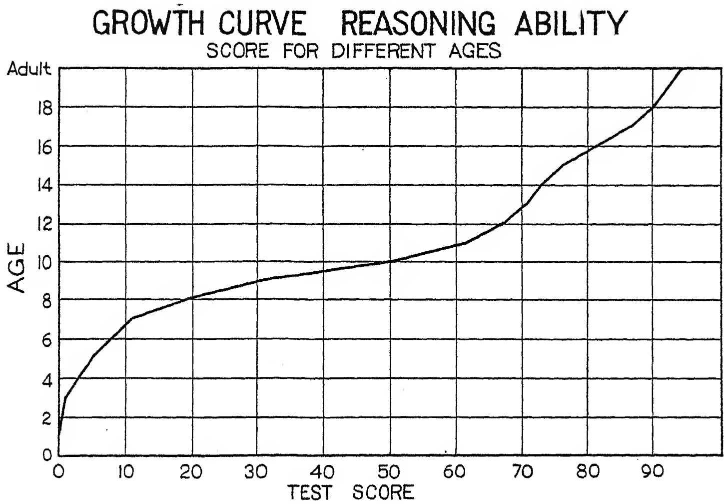


This particular curve is interesting in that it shows a flattening at ages 13 and 14, which is not at all characteristic of growth curves of mental traits, but as the units of measurement, instead of intrinsic ability, could conceivably account for the phenomena the curve does not prove, but merely suggests, that there is a pubertal disturbance. For the purpose of the present statistical treatment no attention need

be paid to the double inflection of the curve.

Rotating the curve through 90° and looking at it in a mirror (as pictured in Chart IV VII) shows its general resemblance to an ogive curve. It was possible in the case of daily temperatures to cumulate scores and obtain ogive curve data. By the reverse process it is possible from the ogive data to obtain the original growth curve. The *growth curve* may be plotted as herewith:

CHART IV VII



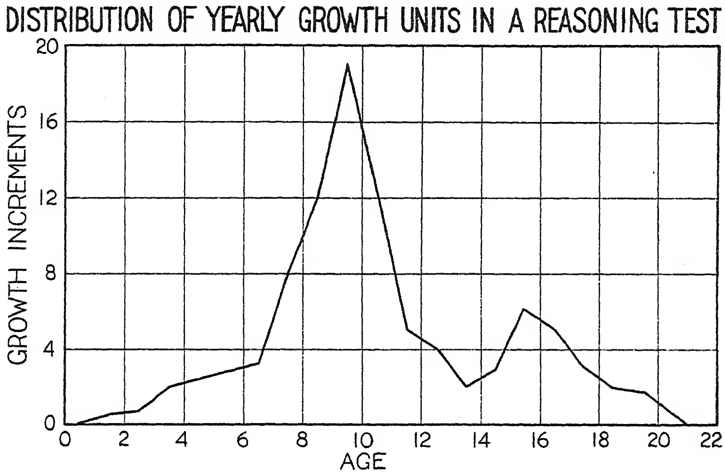
Thinking of the abscissas as sums of increments of reasoning ability and recalling that the graph is for an average individual, whose maximum development or accumulation is to 94 of such increments (i.e., the total population of increments is 94) the graph may be read: At age 7 the individual possesses 11 increments of reasoning ability; at age 10, 50 increments, etc. This may be an awkward way of interpreting growth, but if it is desired to think of growth as a sum of increments it immediately suggests the determination of the increments added during each year of life as follows:

TABLE IV I

AGE	Score	Yearly Growth Increment	Age
0	0		
		0	.5 (from 0-1)
1	0		
		.5-	1.5 (from 1-2)
		.5+	2.5 (from 2-3)
3	1		
		2 -	3.5 (from 3-4)
		2 +	4.5 (from 4-5)
5	5		
		3 -	5.5 (from 5-6)
		3 +	6.5 (from 6-7)
7	11		
		8	7.5 (from 7-8)
8	19		
		12	8.5, etc.
9	31		
		19	9.5
10	50		
		12	10.5
11	62		
		5	11.5
12	67		
		4	12.5
13	71		
		2	13.5
14	73		
		3	14.5
15	76		
		6	15.5
16	82		
		5	16.5
17	87		
		3	17.5
18	90		
		4	? (from 18-adulthood)
Adult	94		
		94	

These growth increments plotted in the form of an ordinary frequency polygon give the graph shown in Chart IV VIII.

CHART IV VIII



The bi-modality of the growth increment curve is of course a consequence of the double inflection of the growth curve. Since the constants of this increment curve (mean, skewness, standard deviation, etc.) can be readily calculated, the curve has certain advantages over the growth curve. It should be a very convenient form in which to present data for purposes of studying variability in rate of growth, variability in price changes, etc. If the modifiability of a function (by education, by promotional sales pressure, by amount of food consumed, etc.) is roughly proportional to the normal rate of change of the function, the growth increment curve is a helpful graphic device.

In dealing with functions in which there is a loss in a given period, e.g., when an individual

weighs less in one year than in the preceding, negative frequencies arise. These need cause no computational trouble if treated strictly algebraically and the negative sign preserved.

Brown and Thomson (1921) have shown that the standard deviations of the class frequencies of such a curve are not given by $\sqrt{Npq}$, the ordinary formula for the standard deviation of the frequency in a class. (See Chapter IX).

SECTION 3. QUANTITATIVE SERIES

When the things observed are such that the measurements taken vary in a negligible manner with change in time when, or place where, taken, we get the usual quantitative or qualitative series. Observations of the height or eye color of children fulfill these conditions if the time interval is so small that growth is negligible. Whether measured on the first of the month, the second, or the fourth, makes no material difference. Again, suppose we give a 30-second speed-of-tapping test. In this function, there is not only considerable variation from day to day but from minute to minute, so that the scores on the first, the second, and the fourth may be quite different. Nevertheless we should, for most purposes, treat such a set of measures as a quantitative series, not because they do not change in time, but because they change in a chance manner only with a change in time. Expressed in another way we can say that the interest we take in these measures is unrelated to changes in time. It not infrequently happens that the same data, depending upon our interest which determines the aspect of it that we study, is now of one type of series and now of another. When time and place are insignificant per se, or for our purposes are such, they are eliminated as axes or dimensions in the graphic portrayal

of the data. There remain, as dimensions for the chart, the value of the trait or character measured and the frequency with which the distinguishable values occur.

The histogram and the frequency polygon: For illustration let us be concerned not with the specific dates connected with the maximum daily temperatures given in Table IV B, but only with the frequencies of occurrence of the different temperatures. We now have a quantitative series. The data are made into a frequency table, as given in the first column of Table IV J, prior to plotting as a *histogram*, Chart IV IX, or *frequency polygon*, Chart IV X.

CHART IV IX

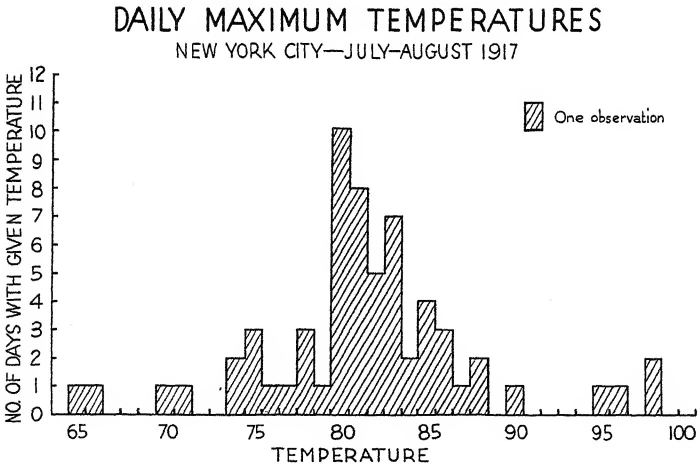
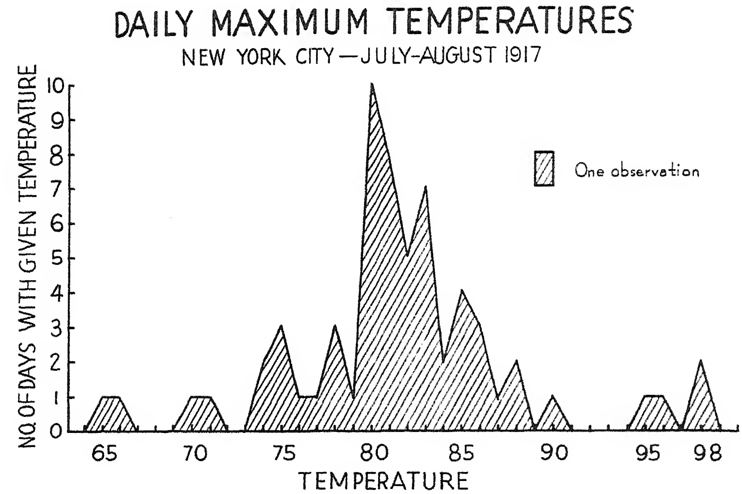


CHART IV X



Either the histogram or the frequency polygon are satisfactory devices for presenting quantitative series. The area under the histogram accurately represents the number of observations, 62 in this case. A little simple geometry shows that this is also the case with the frequency polygon. A somewhat greater idea of continuity is given by the frequency polygon than by the histogram. When drawing the histogram, it is convenient so to record the abscissa scale that the rules of the paper used correspond to boundaries of class intervals. For example, 64.5 and 65.5 are the lower and upper boundaries of the interval within which lie temperatures recorded as 65°, *which value is the mid point of the interval and is called its class index.* The *class limits* are 64.5 and 65.5, and the *class*

interval, designated by the letter *i*, is 1°.

The reader will note that "interval" or "class interval" is used in a double sense. Its most usual meaning is the distance covered by a class (usually the same for all classes), that is, it equals the upper class limit minus the lower class limit. For the data in the 5° column of Table IV J, the interval as thus defined is 5° throughout. "Class interval" may also define some specific class by stating the upper and lower limits. Thus, the second class interval in the 5° column of Table IV J includes values between 67.5° and 72.5°.

TABLE IV J

FREQUENCY DISTRIBUTIONS OF DAILY MAXIMUM TEMPERATURES
NEW YORK CITY, JULY - AUGUST, 1917

Temperatures	Frequencies for Various Groupings												
	i=1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	13°	20°	41°
62.5						1							
63.							2		2				
63.5													
64.													
64.5		1								2			
65.	1				2						4		
65.5				2									
66.	1		2										
66.5		1											
67.	0												
67.5								4					
68.	0												
68.5		0				3							
69.	0		1										
69.5				2								15	
70.	1				2		2						
70.5		2											
71.	1												
71.5													

TABLE IV J (CONTINUED)

FREQUENCY DISTRIBUTIONS OF DAILY MAXIMUM TEMPERATURES
NEW YORK CITY, JULY - AUGUST, 1917

Temperatures	Frequencies for Various Groupings												
	i=1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	13°	20°	41°
72	0		1						8				
72.5		0											
73	0												
73.5				5									
74.	2												
74.5		5				7				13			
75.	3		6		7								
75.5								11					
76.	1												
76.5		2											
77.	1						21						
77.5				6									
78.	3		5								43		
78.5		4											
79.	1												
79.5													
80.	10				27								
80.5		18				34							
81	8		23						41				
81.5				30									
82.	5												62
82.5		12											
83.	7												
83.5								40					
84.	2		13				30						
84.5		6								42			
85.	4				17								
85.5				10									
86.	3												
86.5		4				12							
87.	1		6										
87.5													
88.	2												
88.5		2											

TABLE IV J (CONTINUED)

FREQUENCY DISTRIBUTIONS OF DAILY MAXIMUM TEMPERATURES NEW YORK CITY, JULY - AUGUST, 1917

Temper-	Frequencies for Various Groupings												
atures	i=1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	13°	20°	41°
88.5		2											
89.	0												
89.5				3								47	
90.	1		1		3				7				
90.5		1											
91.	0						3				15		
91.5								4					
92.	0												
92.5		0				2							
93.	0		0										
93.5				1									
94.	0												
94.5		1								5			
95.	1				2								
95.5													
96.	1		2										
96.5		1											
97.	0												
97.5				3									
98.	2						4						
98.5		2				3							
99.			2						4				
99.5								3					
100.					2								

The exact meaning of a recorded score: It has become somewhat customary in educational fields to speak of a child as solving 10 problems in a speed test, meaning thereby that 10 problems were solved and the 11th started but not finished when time was called. In plotting the distribution of scores the designating number,

10, has been placed at the beginning of the interval. No objection should be made to this were the numerical computations in harmony with this procedure, but very generally such scores have been treated as exactly 10.0 in calculating arithmetical averages, with the result that the curve and the constants computed from the data do not agree. Not uncommonly such scores have been treated as 10.0 scores in calculating means and as 10.5 scores in calculating medians, with the result that a comparison of mean and median scores gives an entirely erroneous impression as to the skewness of the data. This faulty procedure has probably been followed unwittingly, but unfortunately with the sanction of teachers. The following is quoted from page 50 of the *Second Year Book*, Division of Educational Research, Los Angeles, July, 1919:

"LESSON SIX—THE ARITHMETIC MEAN

Method of Finding the Mean

No. Problems	No. Pupils	
12	3	3 X 12 = 36
11	5	5 X 11 = 55
10	7	7 X 10 = 70
9	4	4 X 9 = 36
8	2	2 X 8 = 16
		21 213

213 divided by 21 equals 10.14, the mean. The median in this distribution would be 10.64."

In this lesson problem the mean is in error if 12 stands for measures in the interval 12.0 to 13.0 and the median (see Formula [4:01]) is in error if it implies the interval 11.5 to 12.5. The error here cited probably grew out of an error in labeling a distribution. Uniformity is needed, and it would be in harmony with well-nigh uni-

versal procedure in the physical and biological fields to consider a score of 10 as being also a class index, or mid-point of an interval. Should this lower the grade of a few million school children by one-half a point no harm would be done, for all relative positions are maintained.

Certain students have objected to this procedure when the score is zero. Situations exist wherein a negative score is impossible so that a zero class (interval according to the rule −.5 to .5) would actually have all of its frequency in the region .0 to .5. Though the writer has never run across an experimental situation in which the matter was important, should it be so a different procedure, fully explained, as to class indexes and class limits, should be followed. The writer believes that the use of unequal intervals carries with it such annoyance in harmonizing computational procedures leading to the mean with those leading to percentiles, and of both of these with graphic procedures as to warrant the general practive of treating every score (including 0) as a class index.

Throughout this text a score, no matter how derived originally, is uniformly to be interpreted as a class index or mid-point of an interval extending from half an interval below to half an interval above.

A situation of great importance involving the meaning of a score arises in connection with chronological age. A practice which is not universal, but so common as to demand recognition, is to speak of individuals of various age groups, 8-year-olds, 9-year-olds, upon the basis of the age at last birthday, and not of age at nearest birthday. Thus a child 8 years, 11 months, and 29 days will be called an 8-year-old. If this practice were universal, we could unequivocally call the class index of 8-year-olds 8.5, of 9-year-olds 9.5, etc., and we should then regard

a child's age, if it is merely stated that he is an 8-year-old, as 8.5. The lack of universality of a single practice in this important matter necessitates that each author specify in detail the exact meaning of the classes that he gives in any age classification. We shall here follow the practice of considering "8-year-olds" to consist of all those lying between the limits 8.00 and 9.00, which provides a class with mid-point 8.50.

When drawing a frequency polygon it is convenient so to place the abscissa scale that class indexes coincide with rulings of the paper used. The number of days shown in Table IV J as having a maximum temperature of 64° was zero. Accordingly the point (64, 0) is an essential point on the curve. The curve should be definitely terminated at this point and not at such a point as (64.5, 0). *For both the histogram and the frequency polygon it is desirable that there be at least one interval between the bounding rules of the chart and the initial and final values plotted.* Frequently it is desirable to shade the histogram so that the bounds of the curve will not be confused with the rulings of the paper used. Such shading is hardly necessary in the case of the frequency polygon.

The curves as drawn show 11 *modes*, or temperature *values for which the frequency is greater than for immediately smaller or larger values*, as follows: 65.5, 70.5, 75, 78, 80, 83, 85, 88, 90, 95.5, 98. If these 62 days are taken as a sample of July-August days in general, and such would typically be the case as one would be interested in these data as evidence of future phenomena, it is not to be expected that all of these modes are trustworthy, which would mean that one would expect the same modes to occur in 1918, 1919, etc. These sundry modes are features of the sample and are due to chance, or unknown causes, and are not

to be taken as descriptive of the population, the general distribution, or average distribution of maximum daily temperatures if for many years July-August records are combined. Probably the major mode, frequently called "the mode,"—that in the neighborhood of 81°, is a feature that exists in the population and is found here, somewhat distorted, in the sample. Let us now group the data of the first column of Table IV J into larger and larger classes, resulting in the several columns of Table IV J, and note what happens to the sundry modes.

Grouping and labeling classes: The frequencies in the various columns of Table IV J are written opposite their respective class indexes. Note that when *the interval is odd,* 1, 3, 5, etc., *the class index is in each instance divisible by the interval,* and the class limits are fractional, being one-half the size of the interval below and above the class index. Thus, for the interval of 3, the first class has a class index of 66 (which is divisible by 3), a lower limit of 64.5 and an upper limit of 67.5. It is desirable that the class limits be fractional for then, if the original data are recorded to the nearest units, no value falls exactly upon a class limit and there is never ambiguity as to which class to assign a value. *Classes may be unambiguously labeled in any one of the three* ways, as illustrated in Table IV K, for an interval of 3.

The first method is *by recording class indexes* and is frequently the simplest and neatest in appearance. The second method is *by recording class limits* and should be employed if the size of the class interval is not constant throughout the entire range of the data. The third method is *by recording the inclusive integral values* and is very satisfactory when making out tally

TABLE IV K

THE LABELING OF CLASSES
(i = 3)

	x	x	x
First class	66	64.5 - 67.5	65 - 67
Second class	69	67.5 - 70.5	68 - 70
Third class	72	70.5 - 73.5	71 - 73
etc.	etc.	etc.	etc.

sheets from original data. If labeling is by this third method a value such as 70 is immediately seen to lie in the second class, while if labeling is by class indexes one must first note that the value 70 lies closer to 69 than to any other class index before it can be assigned to its proper class.

If the size of the interval is even it is impossible to have an integral class index and a fractional class limit at one and the same time. Of these two the fractional class limit is the more important. Since *when the interval is even* the fractional class index cannot be exactly divisible by the interval *we shall always so select the class limits as to make the lowest integral value in a class divisible by the interval.* Thus, if i = 4, the first class is to be consistent with the following labeling, 65.5, or 63.5-67.5, or 64-67 (the 64 being divisible by 4).

This procedure is to be followed rigorously, including the common and important case when i = 10. Here the classes run as follows: 60-69, 70-79, etc., and the class indexes are 64.5, 74.5, etc.,—not 65, 75, etc.,—and the class limits are 59.5, 69.5, etc.,—not 60, 70, etc. If the original data are dollar and cent magnitudes, but not amounts in fractional parts of a cent, then the classes could be: 60.00-69.99, 70.00-79.99, etc., having class indexes 64.995, 74.995, etc., and class limits 59.995, 69.995, etc. It is to

be noted that even in this case 65, 75, etc., are not the class indexes.

In spite of all efforts to employ classes so that boundary values will not be found in the original data, it will sometimes happen that this cannot be done, so we must adopt *a rule for the allocation of boundary values.* We do not wish always to put the boundary value in the class above, or always in the class below, for either of these procedures will introduce a systematic error. To avoid this, we *adopt the rule of assigning a boundary value to the neighboring value that is even.* For example, the value 66.5 will be called 66, and not 67, and thus it is thrown in the lower class. The value 73.5 is to be called 74, and not 73, and is thus thrown into the higher class. This rule is commonly followed by astronomers and workers in the physical sciences. It is a serviceable one for general adoption, though it does not meet all difficulties. If boundaries of classification can be so arranged that the rule is not called into play, they should be so arranged.

Effect of grouping upon modes: Following this diversion upon grouping and labeling we will return to the question of the effect of grouping upon modes. An examination of the frequencies in the column for which $i = 2$ of Table IV J shows that there are five modes at temperatures as follows: 65.5, 70.5, 74.5, 80.5, 98.5, With a little coarser grouping, $i = 3$, there are three modes, 66, 81, 97.5. With the grouping $i = 4$ there are two modes, 81.5 and 97.5. With the grouping $i = 5$ there is but one mode, 80. With the next coarser grouping, $i = 6$, it would be found that two modes appear. However, we may say that in general the coarser the grouping the greater the tendency to eliminate minor modes.

The coarseness of grouping systematically affects the height of the curve at the mode, as shown in Table IV L herewith.

TABLE IV L

MODAL FREQUENCY WITH VARIOUS GROUPINGS

Size of Interval	1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	13°	20°	41°
Frequency at Mode	10	9	$7\frac{2}{3}$	$7\frac{1}{2}$	$5\frac{2}{5}$	$5\frac{2}{3}$	$4\frac{2}{7}$	5	$4\frac{5}{9}$	$4\frac{2}{10}$	$3\frac{4}{13}$	$2\frac{7}{20}$	$1\frac{21}{41}$

For the purposes of graphic portrayal we should follow the principle of grouping coarsely enough to suppress such secondary modes as are probably due to chance, but not such as are probably evidence of secondary modes existing in the population. In short, we should *aim so to group as to suppress the idiosyncrasies of the sample while preserving the verities of the population.* The precise answer to the question "how coarse to group to accomplish this result?" depends upon the nature of the distribution as well as the size of the sample. Since many distributions are quite similar to the normal distribution in the neighborhood of their major modes, an answer strictly applicable to samples drawn from normal populations may be quite serviceable for many other situations as well, and particularly so if concern is as to the location of the major mode.

Since a graphic portrayal that is misleading even as to the location of the major mode will be more so as to the existence and location of minor modes, we surely should adopt a sufficiently coarse grouping that there is probably not a misjudgment as to the interval in which the population major mode lies.

The detailed steps involved in the solution of this problem and leading to Table IV M are given in Chapter XIII, Sections 10 and 11. The

reader is advised to increase, generally very slightly, the number of classes he employs if his data are leptokurtic, i.e., show frequencies in excess of those normally expected in the extreme tail regions and at the mode.

TABLE IV M

THE NUMBER OF CLASSES TO USE FOR GRAPHIC PORTRAYAL OF SAMPLES OF SIZE N DRAWN FROM A NORMAL POPULATION

N	No. of Classes
4-5	2
6-8	3
9-14	4
15-21	5
22-32	6
33-46	7
47-64	8
65-89	9
90-117	10
118-153	11
154-192	12
193-255	13
256-315	14

Applying the information given in Table IV M, we see that we should employ 8+ classes, to present the distribution of the 62 maximum daily temperatures. The range of temperatures is (98-65+1) or 34°. Dividing 34 by 8 gives 4.25, which to the nearest integer is 4, as the preferred size of interval. A plot with this interval is shown in Chart IV XI. The reader is asked to fix firmly in mind that *a grouping as coarse as this is recommended for graphic portrayal* only, and that a finer grouping, $i=12$, as explained in Chapter VII, *is desirable for statistical computations* leading to measures of central tendency, of variability, and of correlation.

CHART IV XI

DAILY MAXIMUM TEMPERATURES

NEW YORK CITY—JULY-AUGUST 1917

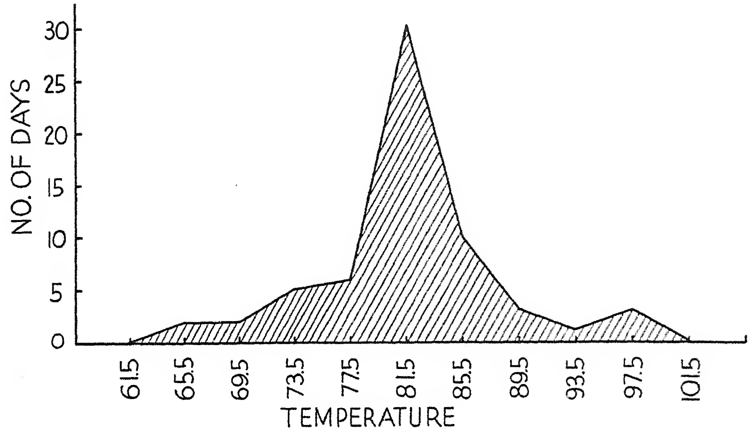


Chart IV XI shows a decided major mode in the neighborhood of 82° and a minor mode in the neighborhood of 98°. The grouping was such that the existence of a major mode in the neighborhood of 82° can be trusted, but the class frequencies in the neighborhood of 98° are so small that the existence of a mode in this region is to be doubted. In other words, the grouping employed was designed to give a certain confidence with reference to the location to within about a half an interval of a major mode only. The more precise method for locating this mode is given in Chapter VII, Section 3. A considerably coarser grouping would be necessary before a much smaller minor mode found in the sample can be assumed to be a feature of the population.

Frequency polygon of a skewed distribution: As an illustration of the tendency for a distribution of ratios to be skewed, and as an important problem in plotting, we will study the data of Table IV N, drawn from Mitchell, as quoted by Secrist (1917).

TABLE IV N

DISTRIBUTION OF 5578 CASES OF CHANGE IN THE WHOLESALE PRICES OF COMMODITIES FROM ONE YEAR TO THE NEXT

PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR (FALLING PRICES)	NUMBER OF CASES	PER CENT OF CHANCE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR (RISING PRICES)	NUMBER OF CASES
54-55.9	1	4- 5.9	356
—	—	6- 7.9	261
50-51.9	1	8- 9.9	237
48-49.9	1	10-11.9	167
46-47.9	1	12-13.9	115
44-45.9	2	14-15.9	106
42-43.9	4	16-17.9	102
40-41.9	5	18-19.9	73
38-39.9	5	20-21.9	65
36-37.9	7	22-23.9	45
34-35.9	10	24-25.9	47
32-33.9	7	26-27.9	29
30-31.9	16	28-29.9	30
28-29.9	27	30-31.9	22
26-27.9	17	32-33.9	17
24-25.9	32	34-35.9	18
22-23.9	39	36-37.9	11
20-21.9	45	38-39.9	17
18-19.9	71	40-41.9	14
16-17.9	76	42-43.9	6
14-15.9	107	44-45.9	10
12-13.9	120	46-47.9	11
10-11.9	173	48-49.9	5
8- 9.9	200	50-51.9	1
6- 7.9	238	52-53.9	4
4- 5.9	329	54-55.9	3
2- 3.9	375	56-57.9	1
UNDER 2	405	58-59.9	6
NO CHANGE	697	60-61.9	4
(RISING PRICES)		—	—
UNDER 2	410		
2- 3.9	355		

TABLE IV N (CONTINUED)

DISTRIBUTION OF 5578 CASES OF CHANGE IN THE WHOLESALE PRICES OF COMMODITIES FROM ONE YEAR TO THE NEXT

PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR (RISING PRICES)	NUMBER OF CASES	PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR (RISING PRICES)	NUMBER OF CASES
—	—	82-83.9	1
66-67.9	4	84-85.9	1
68-69.9	3	86-87.9	1
70-71.9	1	—	—
72-73.9	4	100-101.9	1
74-75.9	1	102-103.9	1
—	—		
80-81.9	1		5,578

It will be noticed that the class intervals extend over ranges of two units, e.g., there are five class intervals in covering a rise in prices from 10 per cent to (but not including) 20 per cent. With no direction to the contrary it is to be presumed that the class designated in the table by "54 - 55.9" includes all measures with values between the limits 53.95 and 55.95; that the next class includes measures between 51.95 and 53.95; etc. This is to say that presumably the data have been recorded to but one decimal place so that such measures as 53.86 and 53.92 are called 53.9 and a measure such as 53.96 is recorded as 54.0. If the recorder encountered a measure 53.95, he had to decide arbitrarily whether it would be called 53.9 or 54.0. For the data in hand it is not definitely known how such a case would have been decided, but we shall assume that the preferred and customary procedure was followed.

If the class intervals run in order from 53.95 to 55.95, 51.95 to 53.95, . . . 1.95 to 3.95, it is found that the next frequency, in order to

extend over the same range, would be from -.05 to 1.95, i.e., from an increase in price of .05 per cent to a decrease of 1.95 per cent. This, however, cannot be the case, as a very large frequency, 697, is recorded for "no change." The way the data are recorded would suggest a class interval corresponding to "no change," but this cannot be so as the intervals on either side preempt the space. The "no change" cases presumably fall at a point and not at variable points within an interval. In plotting the data, therefore, the "no change" interval must be squeezed out and its frequency, 697, distributed between the neighboring classes. We will assign 348 to the "under 2—Falling prices" interval, and the remainder, 349, to the "under 2—Rising prices" interval. There still is a slight discrepancy (.05) in the ranges of these two middle intervals, but as it cannot be positively accounted for without recourse to the original data it is passed over.

For convenience in tabulation and plotting we will consider the first class interval to extend from 54.00 to 56.00 and to have its mid-point or class index 55.00, the second a mid-point at 53.00, etc., and the frequencies as before.

The variable in Table IV N has a range of values from approximately -55.95 to 103.95 or 159.90. The number of cases is 5,578. Reference to Chapter XIII, Table XIII H, suggests that we should have about 33 plotted points, which corresponds to an interval of 4.8. However, the data are very leptokurtic, so that an interval of this size will be coarser than would be desirable in the neighborhood of the mode and finer than desirable in the tail regions. As we are limited in our options to 2, 4, 6, etc., we shall choose an interval of 4, tabulating the data as shown in Table IV O.

TABLE IV O

PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR	NUMBER OF CASES	CLASS INTERVAL OF 4 PER CENT		PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR	NUMBER OF CASES	CLASS INTERVAL OF 4 PER CENT	
		PER CENT OF CHANGE	NUMBER OF CASES			PER CENT OF CHANGE	NUMBER OF CASES
		- 56	1	- 13	120		
- 55	1					- 12	293
—	—			- 11	173		
		- 52	1	- 9	200		
- 51	1					- 8	438
- 49	1			- 7	238		
		- 48	2	- 5	329		
- 47	1					- 4	704
- 45	2			- 3	375		
		- 44	6	- 1	753		
- 43	4					0	1512
- 41	5			1	759		
		- 40	10	3	355		
- 39	5					4	711
- 37	7			5	356		
		- 36	17	7	261		
- 35	10					8	498
- 33	7			9	237		
		- 32	23	11	167		
- 31	16					12	282
- 29	27			13	115		
		- 28	44	15	106		
- 27	17					16	208
- 25	32			17	102		
		- 24	71	19	73		
- 23	39					20	138
- 21	45			21	65		
		- 20	116	23	45		
- 19	71					24	92
- 17	76			25	47		
		- 16	183	27	29		
- 15	107					28	59

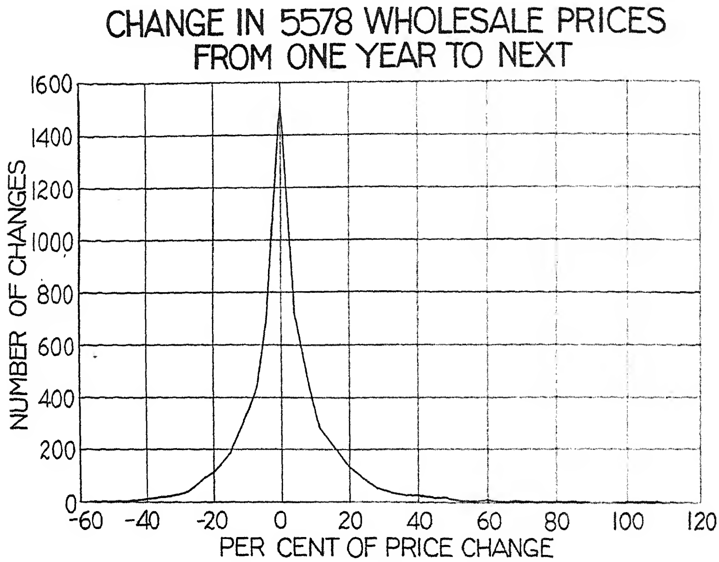
TABLE IV O (CONTINUED)

PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR	NUMBER OF CASES	CLASS INTERVAL OF 4 PER CENT: PER CENT OF CHANGE	CLASS INTERVAL OF 4 PER CENT: NUMBER OF CASES	PER CENT OF CHANGE FROM THE AVERAGE PRICE OF THE PRECEDING YEAR	NUMBER OF CASES	CLASS INTERVAL OF 4 PER CENT: PER CENT OF CHANGE	CLASS INTERVAL OF 4 PER CENT: NUMBER OF CASES
29	30					64	0
31	22			67	4		
		32	39			68	7
33	17			69	3		
35	18			71	1		
		36	29			72	5
37	11			73	4		
39	17			75	1		
		40	31	—	—		
41	14					76	1
43	6					80	1
		44	16	81	1		
45	10			83	1		
47	11					84	2
		48	16	85	1		
49	5			87	1		
51	1					88	1
		52	5	—	—		
53	4					92	0
55	3					96	0
		56	4			100	1
57	1			101	1		
59	6			103	1		
		60	10			104	1
61	4						
—	—				5,578		5,578

The frequency polygon, Chart IV XII, seems better suited to the data in hand, as it gives the impression of a more pronounced mode than would a histogram and in this case, judging by the number of zero change items, this feature

should not be suppressed.

CHART IV XII



Five ways of drawing a graph: There are three common ways of making a graph: (*a*) by erecting rectangles upon the appropriate base intervals; (*b*) by connecting plotted points by means of straight lines; and (c) by drawing a smooth curve through or near all the points which fits the data as nearly as can be determined visually. For quantitative data (*a*) and (*b*) result in the histogram and frequency polygon respectively. A fourth though less common way (*d*) is to plot from smoothed data; and a fifth (e) is the advanced method of mathematically determining the equation of the curve which best fits the data and plotting the same. Methods (*a*), (*b*), and (*e*), and usually (*d*), preserve areas, i.e., the total area under the curve is equal to the number of cases in the sample. Method (e) also preserves other important features.

In using method (c) there should be a definite attempt to preserve areas; that is, if the curve as drawn lies above any point it should lie below some other, or, more accurately, the sum of the vertical distances which it lies above points in the actual distribution should equal the sum of the distances which it lies below other points. In drawing a free curve for such a distribution as that of incomes in the U.S., the preservation of the total area is difficult to insure, but for maximum temperatures, Table IV J and Chart IV X, it can be accomplished with fair accuracy and little trouble. The personal element which enters into method (c) generally makes it inadvisable for published work; but for original, hasty and personal research it may well be used frequently.

The cumulative frequency curve, or ogive, or percentile graph: When the frequencies in the successive classes of an ordinary unimodal frequency distribution are cumulated, interval by interval, beginning at either end, there result frequencies having values of the variate less than, or greater than, the successive limits of the intervals of the frequency distribution. The curve plotted from these is variously called a "cumulative frequency graph," an "ogive," or, in the important case where percentage frequencies are used, a "percentile graph," or "percentile curve."

The temperature data, as given in the first two columns of Table IV J, may be used to illustrate such a graph. From the column of Table IV J giving the finest grouping available, we immediately obtain Table IV P.

The student should carefully note the exact boundary values which apply to the values given in column 3. Table IV J states that the maximum temperature for the coldest day was 65°. As the grouping interval is 1° this asserts that for this day the maximum temperature was greater

TABLE IV P

CUMULATIVE FREQUENCIES OF DAILY MAXIMUM TEMPERATURES
New York City, July-August, 1917

TEMPERATURES	NO. OF DAYS HAVING LOWER MAXIMUM TEMPERATURES	PROPORTION OF DAYS HAVING LOWER MAXIMUM TEMPERATURES
63.5	0	.000
64.5	0	.000
65.5	1	.016
66.5	2	.032
67.5	2	.032
68.5	2	.032
69.5	2	.032
70.5	3	.048
71.5	4	.065
72.5	4	.065
73.5	4	.065
74.5	6	.097
75.5	9	.145
76.5	10	.161
77.5	11	.177
78.5	14	.226
79.5	15	.242
80.5	25	.403
81.5	33	.532
82.5	38	.613
83.5	45	.726
84.5	47	.758
85.5	51	.823
86.5	54	.871
87.5	55	.887
88.5	57	.919
89.5	57	.919
90.5	58	.935
91.5	58	.935
92.5	58	.935
93.5	58	.935
94.5	58	.935
95.5	59	.952
96.5	60	.968
97.5	60	.968
98.5	62	1.000
99.5	62	1.000
etc.	62	1.000

than 64.5° and less than 65.5°. We therefore can assert, as is done in column two of Table IV P, that there was one day with a maximum temperature less than 65.5°. Of course it is incorrect to assert that one day had a maximum temperature less than 65°. Unfortunately this type of error is common and the student is advised to compute points on published percentile charts, if the issue is important and if the basic data are given so that such computation is possible.

It is also to be noted that the information given by the fourth row, that "2 days had a maximum temperature less than 66.5," is just as true and neither more nor less true than that given by the fifth, sixth, or seventh rows, that "2 days had a maximum temperature less than 67.5, 68.5, and 69.5," respectively. Since these are all equally sound we can secure a certain desirable smoothing and a simplicity of statement if, prior to plotting, we replace these four statements by the single statement, "2 days had a maximum temperature less than 68.0," the 68.0 being the average of 66.5 and 69.5. This is recommended procedure wherever multiple statements occur, except in the case of statements referring to terminal situations.

At the lower end we have "zero days had a maximum temperature less than 64.5." Also "less than 63.5." Also "less than 62.5." Etc., without limit. Since the series 64.5, 63.5, 62.5, etc., continues without limit, the average of the smallest and the largest is indefinite. Therefore we will not plot it and we will not assume that we have any trustworthy information from the sample, of what is the temperature such that in the population "zero days will have a less maximum temperature." In other words, *the zero percentile (also the 100) is completely indeterminate from any information given by the sample.* We accordingly will not plot the zero percentile (nor the 100

percentile). In fact, the lowest percentile to be plotted is that given by the proportion of cases in the smallest interval having one or more cases in it. This proportion, in this problem, is 1/62, or .016, so that the lowest computable percentage is the 1.6 percentile. The highest computable percentile is established by one minus the proportion of cases in the highest interval having one or more cases in it. In this problem this proportion is 1-2/62, or .968), so that 96.8 is the highest computable percentile.

Another way of expressing these facts is to say that from this sample of 62 days percentiles less than 1.6 and greater than 96.8 have infinite standard errors, while percentiles between these two values have finite standard errors, though as will be shown later the presumptive errors in the computed percentiles near 1.6 and 96.8 will be very large.

If we smooth as suggested we obtain the following table:

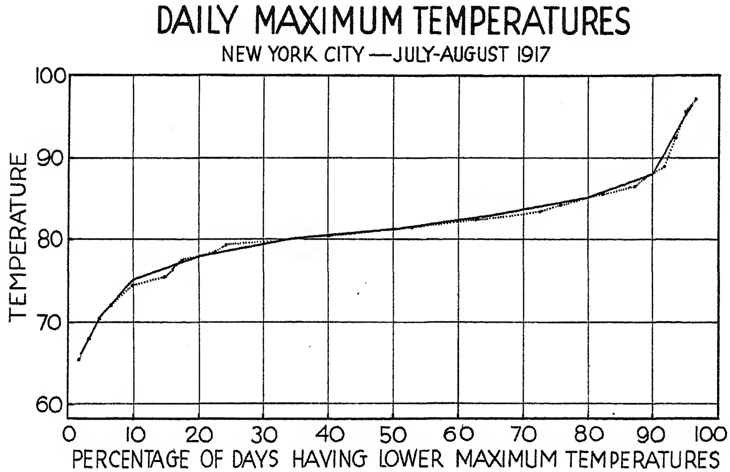
TABLE IV Q

CUMULATIVE FREQUENCIES OF DAILY MAXIMUM TEMPERATURES
NEW YORK CITY, JULY-AUGUST, 1917

Temperatures	Proportion of days having lower maximum temperatures	Temperatures	Proportion of days having lower maximum temperatures
?	.000	81.5	.532
65.5	.016	82.5	.613
68.0	.032	83.5	.726
70.5	.048	84.5	.758
72.5	.065	85.5	.823
74.5	.097	86.5	.871
75.5	.145	87.5	.887
76.5	.161	89.0	.919
77.5	.171	92.5	.935
78.5	.226	95.5	.952
79.5	.242	97.0	.968
80.5	.403	?	1.000

The plot of these data is shown by the dot line curve of Chart IV XIII. Special care is called for in the labeling of the axes. The terms "ogive" and "percentile" are technical terms and somewhat undesirable for use in the title to the Chart. They are quite unsatisfactory in connection with axes labels.

CHART IV XIII



The values for the chart as shown have been laborious to compute, as there are many of them, and annoying to plot, as they are not integral values. It is accordingly economical and not misleading to calculate a small number of selected percentiles and plot these values only. We have earlier noted that for a sample of 62 eight classes are sufficient for graphic portrayal. Some number of computed percentiles, say not less than eight, should suffice for a satisfactory ogive. Let us therefore compute the following percentiles: 2, 5, 10, 20, 35, 50, 65, 80, 90, 95, 96, and plot the ogive from these. The values 2 and 96 have been chosen because they

are the smallest and the largest integral percentiles calculable from the data.

The computation of percentiles: The computation follows formula [4:01] which is simple to use and is self-explanatory as soon as the meaning of the symbols is grasped.

N = the number of cases in the sample.

P_p = the percentile in question (if p = .10 then $P_{.10}$ = the tenth percentile).

p = the proportion of cases having smaller values than P_p.

pN = the number of cases having smaller values than P_p. If exactly pN cases lie below some class limit, that limit $= P_p$. (Note, however, $P_{.05}$ example following.)

Determine the class wherein the pN measure lies and let

v_p = the value of the lower limit of this class.

f_p = the frequency, or number of cases, in this class.

i_p = the interval, or range, covered by this class.

F_p = the sum of the frequencies in all classes below (i.e., classes with smaller X values than) this class.

Then

$$P_p = v_p + \frac{pN - F_p}{f_p} i_p \qquad \text{Value of a percentile} \qquad [4:01]$$

This formula is convenient for finding an assigned percentile.

A not infrequent problem is to find the percentile value of an assigned score. Let this assigned value be P_p. Then every term in [4:01] except p is known and this may be solved for, thus:

$$p = \frac{F_p + \dfrac{f_p(P_p - v_p)}{i_p}}{N} \quad \text{Proportion of cases falling below } P_p \qquad [4:02]$$

Example *(a)* involving no complications: compute the 10th percentile for the temperature data of column "$i = 1°$" of Table IV J and Table IV P. The computation could also be based upon the frequencies in column "$i = 2°$" with satisfactory results, for the number of classes is then still greater than 12, but still coarser grouping should be avoided.

$p = .10$

$pN = 6.2$

We observe that the 6.2 measure lies in the class with class index 75.0 and that

$v_p = 74.5$

$f_p \quad 3$

$i_p \quad 1$

$F_p \quad 6$

Accordingly

$$P_{.10} = 74.5 + \frac{.10(62) - 6}{3} 1 = 74.57$$

Example *(b)* wherein the class containing the pN measure is neighboring to one or more classes with zero frequency: Compute $P_{.05}$.

$p = .05$

$pN = 3.1$

We observe that the 3.1 measure lies in the class in Table IV J with class index 71 and that the two neighboring classes above have zero frequency. We therefore consider the frequency, 1, found in this class to extend over a stretch equal to that of this class plus half that of the neighboring zero classes, so that

$v_p = 70.5$ and the upper class limit $= 72.5$ giving

$i_p = 2$

$f_p = 1$

$F_p = 3$

Thus

$$P_{.05} = 70.5 + \frac{.05(62) - 3}{1} 2 = 70.70$$

Having the computed values for percentiles conveniently spaced from the smallest, $P_{.02}$, to the largest $P_{.96}$, the percentile curve may quickly be plotted, as shown by the full line curve of Chart IV XIII. Comparison with the dot line curve shows that this curve based upon but 11 values, all integral, is entirely adequate.

The issue of the reliability of percentiles could be appropriately studied after that of variability, as given in Chapter VI, but to impress the inescapable intimacy between a statistic and its error the standard error of a percentile is introduced at this point. The beginning student is not expected to grasp fully the meaning of a standard error at this point. After studying Chapter VI he should re-read this section.

In calculating the standard error of a percentile it is desirable to use an interval as large as that recommended for graphic portrayal and to make the center of the interval coincide as nearly as possible with the percentile in question. We therefore define two new constants, i_p' and f_p'.

i_p' = the interval recommended for graphic portrayal.

f_p' = the frequency in the i_p' interval having a class index as nearly equal to P_p as the original grouping (i.e., the data as first recorded prior to any grouping) of the data permits.

We also let $q = 1 - p$

Then,

$$\sigma_{P_p} = \frac{i_p' \sqrt{Npq}}{f_p'}$$ Standard error of a percentile*.............. [4:03]

In the derivation of this, as of all other standard error formulas, the elements in the right-hand member call for population, or true, values. Since these are unavailable, necessity dictates that sample values be substituted, with the result that the standard error computed does itself have a standard error or element of uncertainty, but as this usually attaches to more remote figures than the first or second, the computed value remains a highly important statistic.

* Derived in Kelley, 1923, Formula [42].

SECTION 4. QUALITATIVE SERIES

The qualitative or categorical is the most primitive of all statistical series. It conveys the minimum of information about the data, informing one simply that there are different classes and different numbers of cases or different values attached to the different classes. If certain temporal, geographic, or quantitative information inherent in series of these types is neglected in studying the data of these types, the series degenerate into qualitative series, for there still remain different categories and attached amounts or frequencies.

There are two important sub-types to qualitative data: (1) in which random sampling of a population has resulted in observed frequencies in a number of categorical classes, and (2) in which an observation of some specific sort taken upon a number of categories shows different values, not numbers of cases, attached to the respective classes. The second column of Table IV R (hypothetical) is of the first type, while the third column is of the second. The distinction between these sub-types is important in

TABLE IV R

VOCATIONAL DISTRIBUTION IN NOTOWN OF WORKERS AND OF CAPITAL LOCALLY INVESTED IN CONNECTION WITH THEIR BUSINESSES OR EMPLOYMENTS

Vocations	Number of Workers	Capital Employed
Farming	80	$180,000
Mining	7	35,000
Transportation	8	10,000
Manufacturing	45	110,000
Selling	20	40,000
Personal Service	10	5,000
Other	30	20,000
	200	$400,000

connection with computational procedures, for the reliability of the frequency in classes obtained as a result of random sampling is known (see [9:02]) but the reliability of amounts is not ascertainable without more information than is commonly available. The reason for this is that each case,—worker employed,—is presumably an independent observation resulting from a process of enumeration, but each amount,—dollar of capital employed—is presumably not an observation which is independent of every other amount. We may not reasonably look upon the $400,000 as a result of 400,000 measures chosen at random from an infinite population. If there exists Sometown, which upon acquaintance is judged to be generally similar to Notown and having 200 workers, we may expect to find, within limits suggested by

$$\sqrt{\frac{200 \times 45 \times 155}{200 \times 200}}\ (= 5.9),$$ the standard deviation of the frequency in a class, that the number of workers engaged in manufacturing will be 45, but we should not expect to find, within limits suggested by

$$\sqrt{\frac{400{,}000 \times 110{,}000 \times 290{,}000}{400{,}000 \times 400{,}000}}\ (= 282),$$

that the amount of capital employed in manufacturing will be $100,000. This difference is of great importance in statistical deduction. It is even important in the reading of a graph.

Customary devices for representing categorical data are the bar graph, Chart IV XIV, *the segmented bar, the circle graph or pie chart* which is a circle with sectors, Chart IV XV, *and pictograms* representing both the number of cases in and the nature of categories, as, e.g., would be done if 80 farmers are represented by, say, 16 pictures of men with hoes, and seven miners by 1.4 men of the same size as the farmers but with miners' caps and lamps, etc. The first three

devices require descriptive lables, but the pictogram is serviceable for non-readers or others more inclined to the comics than the editorial page.

CHART IV XIV

WORKERS AND CAPITAL IN NOTOWN

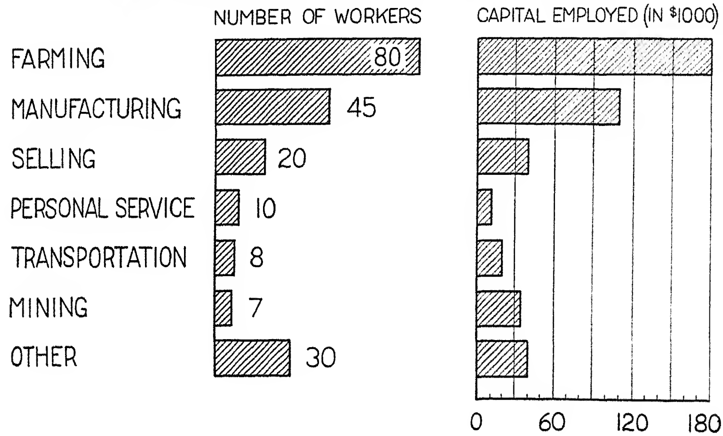
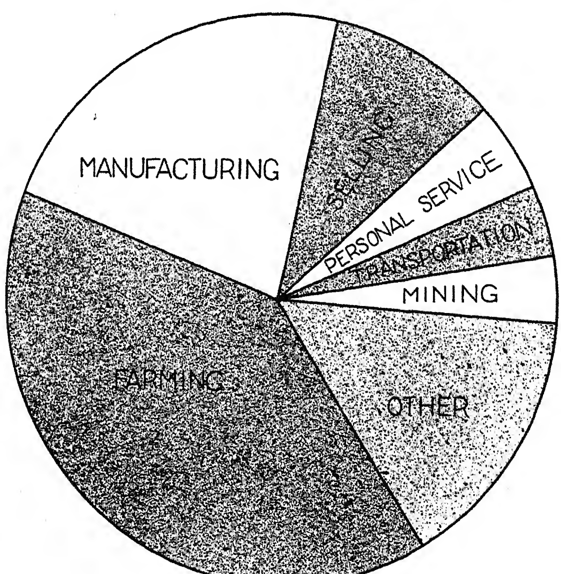


CHART IV XV

DISTRIBUTION OF CAPITAL IN NOTOWN



SECTION 5. GEOGRAPHIC SERIES

The feature of the spatial series not shared by quantitative, qualitative, or temporal series *is that the data are tied to a position in space which* is important for the issues in question and *must not be neglected.* If neglected the series ceases to be spatial and becomes, generally, a quantitative series. This is commonly the case when names instead of geographic positions are given in connection with products of districts, e.g., if the 1941 chicken production for the counties of California are given in a table naming the counties and the number of chickens produced, the reader who does not have a map of California, showing counties, before him or in his mind's eye is presented with nothing but a quantitative series. If this information is presented on a map it is a geographic series, which is the most common sub-type under spatial series. Just as the temporal series for the daily maximum temperatures in July and August in New York City ceased to be temporal when the temperatures were divorced from the particular dates with which originally connected, so a geographic series ceases to be such when the items are so presented that the reader does not attach them to geographic position. For the existence of a spatial series a spatial metric,—one, two, or three dimensions as the case may be,—must be employed.

The one- and the three-dimensional spatial series are unusual but very interesting when relevant. A survey giving for New York City the amounts of dust in the air per cubic foot for different spatial positions, each having a north-south, an east-west, and a distance-above-street-level dimension, is an illustration of a three-dimensional spatial series. Its representation requires a three-dimensional model (or a stereo-

scopic picture) with the dust density records, by density of a swarm of points, by figures, by color, or otherwise, attached to their respective spatial positions. Numerous illustrations can be drawn involving sea-life populations and their density in different locations in the ocean. If we add seasons of the year such a series becomes four-dimensional, with time as the fourth dimension.

The one-dimensional spatial series is uncommon. A precise illustration of such is to be found in a chromosome map. The location of genes in a chromosome is strictly linear. To present such a series we require a line, straight or curved provided it does not cross itself, with the location of the genes located thereon. In this simple case, if the line with the located genes is not given, but only the names of the genes, the spatial series disappears and we have nothing but a qualitative series left.

When we note that debased spatial, quantitative, and temporal series reappear as qualitative series, we are led to ask if qualitative series which are not recognized as degenerate series of the other types would not in fact be seen to be such if fuller information were available. We may also note in this connection *that the statistics, deriving from frequencies and proportions in classes, appropriate, to qualitative series are far more primitive in their nature, though at the same time more universal in their applicability, than the statistics that serve a quantitative series.* Both because of this and because of the loss of information which is common when data are merely thrown into categories, the student should be content with qualitative data and treatment only in case the data cannot be cast into one of the other three types.

It is somewhat difficult to classify as two- or as three-dimensional data distributed over

the surface of the globe. Since the data all lie upon the surface a much more limited type exists than the general three-dimensional spatial series. Certainly if we are concerned with so small a portion of the surface of the sphere that it can approximately be represented by a plane, we have a two-dimensional spatial series. This is the most common situation and we shall think of geographic series as essentially two-dimensional spatial series.

In dealing with data attached to the forty-eight states of the United States we should follow the tabulating practice of the U. S. Census. This is shown in the stub of Table IV S. If we desire to present the population of the states as a geographic series we must represent the population figures upon a map in their appropriate positions, by writing the population figures, by varied cross-hatching, by shading, or by stippling. In these last three cases the greater the density of population the heavier the hatching, shading, or stippling. Shading or stippling generally yield the more accurate visual impression.

The population data just mentioned may be called a quantitative geographical series because different quantities of a single thing are attached to the different geographical positions. Should the chief products of the states be shown on a map,—a steer being drawn on Texas, a shock of wheat on Kansas, etc., we have a qualitative geographic series. This latter is very common in popular portrayal and is serviceable if there is no attempt to convey other than qualitative information. Of course the veriest tyro is aware that such a presentation fails to reveal important quantitative aspects that exist.

The most primitive geographic series merely shows that different regions exist. Such regions are defined by the boundaries given on the map.

TABLE IV S

CERTAIN POPULATION, AREA, AND SCHOOL ATTENDANCE STATISTICS OF THE UNITED STATES, BY STATES, 1940*

DIVISION AND STATE	POPULATION IN 1000'S	AREA IN SQUARE MILES	PERSONS OF AGE 16-20: TOTAL NO. IN 1000'S	PERSONS OF AGE 16-20: ATTENDING SCHOOL (IN 1000'S)	PERSONS OF AGE 16-20: PER CENT
UNITED STATES	131,669	2,977,128	12,278	5,106	41.6
NEW ENGLAND	8,437	63,206	761	329	43.2
Maine	847	31,040	76	31	41.2
New Hampshire	492	9,024	43	17	41.0
Vermont	359	9,278	32	13	41.1
Massachusetts	4,317	7,907	384	175	45.6
Rhode Island	713	1,058	68	24	35.1
Connecticut	1,709	4,899	157	67	42.8
MIDDLE ATLANTIC	27,539	100,496	2,475	1,089	44.0
New York	13,479	47,929	1,135	526	46.3
New Jersey	4,160	7,522	379	155	40.8
Pennsylvania	9,900	45,045	961	409	42.6
EAST NORTH CENTRAL	26,626	245,011	2,357	1,047	44.4
Ohio	6,908	41,122	622	288	46.3
Indiana	3,428	36,205	308	134	43.5
Illinois	7,897	55,947	677	291	43.0
Michigan	5,256	57,022	471	203	43.0
Wisconsin	3,138	54,715	280	131	47.0
WEST NORTH CENTRAL	13,517	510,621	1,235	533	43.2
Minnesota	2,792	80,009	256	113	44.3
Iowa	2,538	55,986	230	98	42.7
Missouri	3,785	69,270	331	126	37.9
North Dakota	642	70,054	65	28	43.2
South Dakota	643	76,536	64	30	46.6
Nebraska	1,316	76,653	123	56	45.2
Kansas	1,801	82,113	166	83	50.0
SOUTH ATLANTIC	17,823	268,431	1,851	619	33.4
Delaware	267	1,978	24	9	38.3
Maryland	1,821	9,887	169	56	32.9
Dist. of Columbia	663	61	51	24	46.9
Virginia	2,678	39,899	280	93	33.1

TABLE IV S (CONTINUED)

CERTAIN POPULATION, AREA, AND SCHOOL ATTENDANCE STATISTICS OF THE UNITED STATES, BY STATES, 1940*

DIVISION AND STATE	POPULATION IN 1000'S	AREA IN SQUARE MILES	PERSONS OF AGE 16-20: TOTAL NO. IN 1000'S	ATTENDING SCHOOL (IN 1000'S)	PER CENT
(CONTINUED)					
SOUTH ATLANTIC					
West Virginia	1,902	24,090	200	73	36.3
North Carolina	3,572	49,142	401	132	32.8
South Carolina	1,900	30,594	224	69	31.0
Georgia	3,124	58,518	330	99	29.8
Florida	1,897	54,262	170	65	37.9
EAST SOUTH CENTRAL	10,778	180,568	1,106	374	33.9
Kentucky	2,846	40,109	287	81	28.4
Tennessee	2,916	41,961	293	100	34.0
Alabama	2,833	51,078	297	107	35.8
Mississippi	2,184	47,420	228	87	37.9
WEST SOUTH CENTRAL	13,065	430,829	1,305	505	38.7
Arkansas	1,949	52,725	201	72	35.9
Louisiana	2,364	45,177	239	82	34.2
Oklahoma	2,336	69,283	236	109	46.3
Texas	6,415	263,644	629	242	38.4
MOUNTAIN	4,150	857,836	395	188	47.5
Montana	559	146,316	52	26	50.0
Idaho	525	82,808	52	26	50.8
Wyoming	251	97,506	24	11	46.6
Colorado	1,123	103,967	102	47	45.9
New Mexico	532	121,511	53	22	41.1
Arizona	499	113,580	47	19	40.8
Utah	550	82,346	57	32	56.7
Nevada	110	109,802	8	4	50.6
PACIFIC	9,733	320,130	793	422	53.2
Washington	1,736	66,977	149	79	53.0
Oregon	1,090	96,350	92	46	50.2
California	6,907	156,803	551	296	53.8

* Abstract of the Sixteenth Census of the U.S., 1940, p. 45 and 76, Population, Second Series, Characteristics of the Population, United States Summary.

To make these regions more easily observed they may be colored differently. It is not necessary to employ as many colors as there are regions, for non-contiguous regions may be colored the same without confusion. Experience, rather than mathematical deduction, shows that if the region to be mapped is the globe or any portion thereof four colors will always suffice to meet the requirement that no two contiguous regions shall be colored the same. In any actual case it may call for trial and retrial to find the proper order of coloring to insure that this minimal number four suffices.

Frequently the map with the recorded phenomena upon it constitutes the end product, or the terminal geographic statistic. However, computational techniques are appropriate to the handling of important issues that commonly arise. When the nearly waste land that is now Gary, Indiana, was established as a steel center, it was because of its strategic geographic location. To be considered was the cost of bringing ore from Superior, coal from southern Illinois, and of distributing finished products to railways and industrial plants in the Middle West and beyond. Taking these and other necessary considerations, there existed some location that would minimize costs and maximize profits. Gary, after a careful survey, was judged to be it. The problems of geographic series may become very complex and they arise in many business and social intercourse situations.

The school superintendent recommending the building of a new school should do so primarily to minimize the distance and the hazards that pupils are subjected to in going from home to school. Of the available locations, undoubtedly some one better accomplishes these things than any other. The basic treatment of this problem requires first of all a plotting of a school

population map, as estimated to exist at a future date, corresponding to one-half the life of the school building to be constructed. Upon this map lines may be drawn representing the hazards, e.g., more lines for railway crossings and boulevards than for residential streets. Then each site may be tested by making computations giving the sum of pupil distances from school and the sum of hazards crossed. If some such combining concepts as "one-half mile of distance is equally disadvantageous to crossing one boulevard" are made, distance debits and hazard debits may be combined and a single figure obtained representing the total debit to be attached to a site. A comparison of sites by this means will disclose that site having minimal distance and hazard debits.

The procedure suggested is closely related to that of determining the center of population. Herewith is a brief table, taken from Read (1939), of centers in the United States of certain populations.

TABLE IV T

THE CENTERS OF POPULATION OF CERTAIN LEARNED GROUPS IN THE UNITED STATES

ORGANIZATION	NO. OF INDIVIDUALS	CENTER OF POPULATION		NEAREST LARGE CITY
		Latitude	Longitude	
College enrollment	885,282	39°19′	84°45′	Cincinnati
Am. Chem. Soc. by sections	17,469	40° 5′	83°49′	Dayton
Am. Historical Society	2,926	39°54′	82° 13′	Columbus
Math. Assn. of America	1,928	39°32′	84° 57′	Cincinnati
Am. Assn. of Petroleum Geologists	2,448	35° 6′	98°12′	Oklahoma City
Am. Psychological Assn.	2,269	40° 15′	83°35′	Columbus
Am. Assn. of University Professors	11,165	39°30′	84°27′	Cincinnati-Dayton
Am. Assn. for the Advancement of Science	17,141	39°41′	84°10′	Dayton

The method of calculation of such centers followed, with minor modification, the general procedure of the United States Census in its computation of the center of the United States population, as described in full on page 20, Volume II, of the 15th Census of the U.S., and as here quoted in abridged form:

"The center of population may be said to represent the center of gravity of the population. If the surface of the United States be considered as a rigid level plane without weight and the population distributed thereon, all individuals being assumed to have equal weight, the point on which this plane would balance would be the center of population. . . .

"In making the computations for the location of the center of population it is necessary to assume that the center is at a certain point. Through this point a parallel and a meridian are drawn, crossing the entire country. . . .

"The product of the population of a given area by its distance from the assumed parallel is called a north or south moment, and the product of the population of the area by its distance from the assumed meridian is called an east or west moment. . . .

"The population of each square degree north and south of the assumed parallel is multiplied by the distance of its center from that parallel; a similar calculation is made for the principal cities; and the sum of the north moments and the sum of the south moments are ascertained. The difference between these two sums, divided by the total population of the country, gives a correction to the latitude. In a similar manner the sum of the east and of the west moments are ascertained and from them the correction in longitude is made."

As a simple center of population problem let us consider the following, which we will call the Problem of the Statisticians' Picnic: Twenty Massachusetts statisticians with residences as given in Table IV U decide to have a picnic and, without prior determination, agree that the site shall be the center of population of the twenty.

TABLE IV U

PICNICKING STATISTICIANS OF MASSACHUSETTS

NAME OF STATISTICIAN	PLACE OF RESIDENCE	NAME OF STATISTICIAN	PLACE OF RESIDENCE
Adams	N. Adams	King	Gloucester
Pitt	Pittsfield	Bridge	Cambridge
Barr	Great Barrington	Bean	Boston
Spring	Springfield	Fern	Boston
Green	Greenfield	Ivy	Boston
Woo	Worcester	Rock	Plymouth
Fitch	Fitchburg	Town	Provincetown
Ames	Amesbury	Ford	New Bedford
Lawrence	Lawrence	Wood	Woods Hole
Lowell	Lowell	Wells	Wellesley

Where is this Mecca? We solve this problem by taking measurements on a map of Massachusetts,

CHART IV XVI

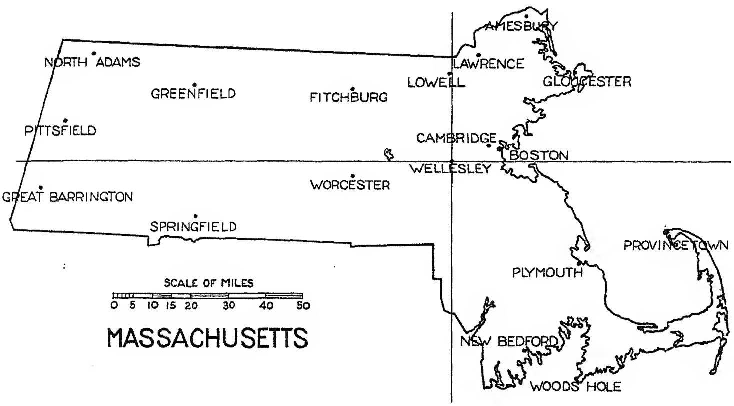


Chart IV XVI. We choose some arbitrary point, say Wellesley, somewhere near the center of population, and draw a meridian and a parallel through it. Assuming a flat surface and using the scale of the map and a pair of calipers, we determine the north-south and the east-west coordinates of each of the twenty residences. The figures of Table IV V result.

TABLE IV V

NORTH-SOUTH AND EAST-WEST COORDINATES OF RESIDENCES FROM WELLESLEY

NAME	N-S DISTANCE IN MILES	E-W DISTANCE IN MILES	$(\text{N-S DISTANCE})^2$	$(\text{E-W DISTANCE})^2$
Adams	28	- 93	784	8649
Pitt	10	-100	100	10000
Barr	- 7	-106	49	11236
Spring	-14	- 66	196	4356
Green	20	$-67\frac{1}{2}$	400	$4556\frac{1}{4}$
Woo	- 3	- 26	9	676
Fitch	19	$-25\frac{1}{2}$	361	$650\frac{1}{4}$
Ames	38	22	1444	484
Lawrence	28	$7\frac{1}{2}$	784	$56\frac{1}{4}$
Lowell	23	$-\frac{1}{2}$	529	$\frac{1}{4}$
King	22	32	484	1024
Bridge	4	$9\frac{1}{2}$	16	$90\frac{1}{4}$
Bean	$3\frac{1}{2}$	$11\frac{1}{2}$	$12\frac{1}{4}$	$132\frac{1}{4}$
Fern	$3\frac{1}{2}$	$11\frac{1}{2}$	$12\frac{1}{4}$	$132\frac{1}{4}$
Ivy	$3\frac{1}{2}$	$11\frac{1}{2}$	$12\frac{1}{4}$	$132\frac{1}{4}$
Rock	$-23\frac{1}{2}$	31	$552\frac{1}{4}$	961
Town	$-16\frac{1}{2}$	56	$272\frac{1}{4}$	3136
Ford	-44	19	1936	361
Wood	$-53\frac{1}{2}$	32	$2862\frac{1}{4}$	1024
Wells	0	0	0	0
Sums	41	-241	$10815\frac{1}{2}$	47657
Means	2.05	- 12.05	540.775	2382.85

$\sqrt{540.775 - (2.05)^2 + 2382.85 - (12.05)^2} =$
$\sqrt{2774.22} = 52.67$

Computing sums and means we find that the center of population of these twenty is 2.05 miles north and 12.05 miles west of Wellesley. Had the faith of the statisticians in their subject been wavering, it would have been revived as they found their rendezvous,—a beautiful little island in the middle of a fine body of water.

The center of population defined in line with the computations made has the important property of minimizing the sum of the squared distances, or the quadratic mean distance. For the picnic problem we find this to be

$\sqrt{540.775 - (2.05)^2 + 2382.85 - (12.05)^2}$, or 52.67 miles.

Should the desire be to minimize the mean, or the median, or the modal distance, other definitions and other computational procedures would be necessary. Concerning the modal distance we may note that if the place of greatest density is chosen the number zero distance from the locus is maximized, but if the frequency of cases at distance x $(x \neq 0)$ is to be maximized, some point (or points) other than this place of densest population will in general constitute the solution.

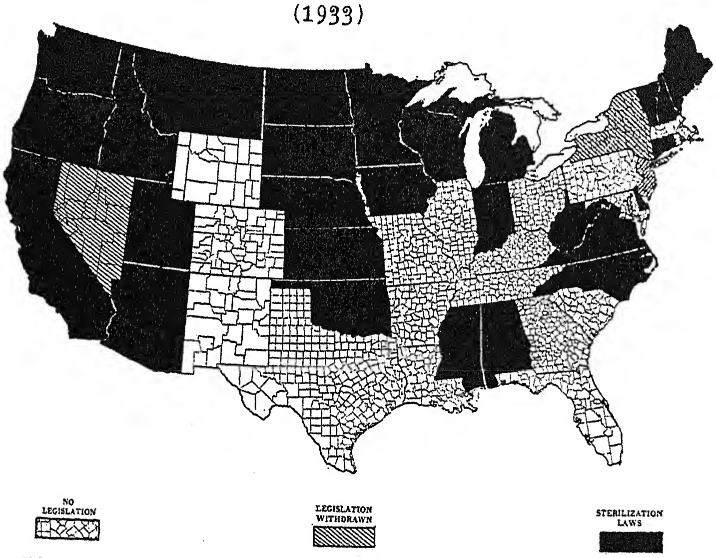
If the geographic phenomena studied are continuous and show moderate rates of change over the region studied, they can be excellently portrayed by means of contour lines, as is done in isothermal and isobaric weather maps and in Coast and Geodetic Survey maps giving land elevations and sea depths. Military maps of terrain illustrate a most precise portrayal of a geographic series.

Chart IV XVII taken, with permission, from *Birth Control Review*, Vol. 17, no. 4, attempts to depict the status of sterilization laws in the United States in 1933. It is not entirely satisfactory for the states shown in black do not constitute a uniform class. There is, of course, wide diversity in the statutes in these several states.

CHART IV XVII

STERILIZATION LEGISLATION IN THE UNITED STATES

(1933)



The mapping on a two-dimensional sheet of so large a portion of the surface of the globe that it cannot be represented satisfactorily by a plane, follows different practices depending upon what feature, or features, it is most important to portray correctly.

The Mercator projection of the surface of the globe is obtained by circumscribing the globe with a thin paper cylinder which touches it at the equator and projecting the points on the globe onto the outside of the cylinder, by using lines

that pass through the center of the globe. This projection has the important property of preserving all compass directions, as is important to the navigator, but areas and distances are not preserved and are, in the arctic regions, greatly exaggerated.

An orthographic projection is obtained by projecting the points of the surface of one-half the globe onto a plane which touches the globe at one point (the center of the map),—the projecting lines being perpendicular to the plane. Two such circle maps are commonly found in an elementary geography to picture the entire global surface, but distances and areas are greatly distorted, except in the neighborhoods of the two centers.

Another mapping designed to serve a single spot is an azimuthal equidistant projection. If Washington were made the center of such a map the rest of the globe would be so presented that all distances from Washington would be correctly indicated by the straight-line distances from Washington. Thus all places 1000 miles away would lie on a circle of indicated radius 1000 miles, having its center at Washington. A dimension at right angles to the distance-from-Washington dimension is more and more exaggerated as the distance from Washington becomes greater, but the fact that all straight lines through Washington correspond to great circles on the globe through Washington is a merit when the issue is distance from Washington. A company engaged in international trade might well desire such a map with its headquarters, or producing plant, as the center.

A gnomonic projection is one constructed with the center of the sphere the vantage point, but when viewed by the reader the viewpoint is not the center, but from points outside the sphere. Only a portion of the globe can be so represented. The merit of this projection is that for the

portion shown the straight line connecting any two points (not merely the center and a second point) corresponds to the great circle on the globe connecting these two points.

There are mappings which can be made of all, or portions of, the globe so as to preserve areas. One such is the Aitoff equal-area projection map. It is elliptical in shape, with the equatorial dimension twice that of the inter-polar dimension. Another is the sinusoidal interrupted projection, which consists of a succession of lens-shaped parts whose touching points lie along the equator.

Though the globe is the unbiased authority for spheric relationships, still these sundry two-dimensional maps do well serve the special problems involving the direction, distance, or area, or a particular point of reference.

SECTION 6. GRAPHIC PORTRAYAL IN THREE DIMENSIONS

The two-dimensional base is serviceable in other than geographic problems. These essentially are situations involving correlated variables, in which the frequency of occurrence varies depending upon the joint values of two variables. If the variables are quantitative we have the common scatter diagram underlying simple correlation. If the data are qualitative we have the common contingency table underlying coefficients of contingency and χ^2 tests of independence of variables. In either of these cases it may be serviceable to present the data graphically by means of a block diagram.

A rectangular parallelopiped, for brevity here called a block, has three dimensions. As it stands upon a flat surface the two dimensions in contact with the surface may represent the two different variables and the height of the block can represent the frequency of the joint occurrence. For accurate portrayal the blocks must

be drawn in perspective and the vantage point from which the blocks are observed must be such that at least a part of each block shall be visible. If some block is very low, i.e., the corresponding frequency small, so that it is a well surrounded by towering neighboring blocks, or at least touching towering neighboring blocks in front of it, the vantage point must be at a high elevation in order to glimpse any portion of this block. However, just as hilly country looks flat when viewed from a high-flying aeroplane, so distinctions in the elevations of blocks are difficult if the vantage point is very high. The requirements of low vantage point and complete visibility are frequently antagonistic. It is accordingly true that many data cannot well be represented by means of a block diagram. If it is possible to envisage every block from a relatively low vantage point, it may be expected that presentation by means of a block diagram will be clear and informative. A very common situation arises where one only of the two variables is qualitative, the other being either temporal or quantitative. In this case the temporal, or quantitative, order must be preserved in the corresponding dimension, but the order in the qualitative variable can be chosen because of lucidity of presentation and for no other reason. The data of Table IV W as represented in Chart IV XVIII, illustrate this case.

The successive steps in the construction of a block diagram: We will illustrate these by means of the data given.

The states are listed in alphabetic order in Table IV W. It is not important to preserve this, and to do so would increase the difficulties of presentation. An arrangement of states based upon the populations in 1930 is suggested, but generally a preferable one, to avoid hidden

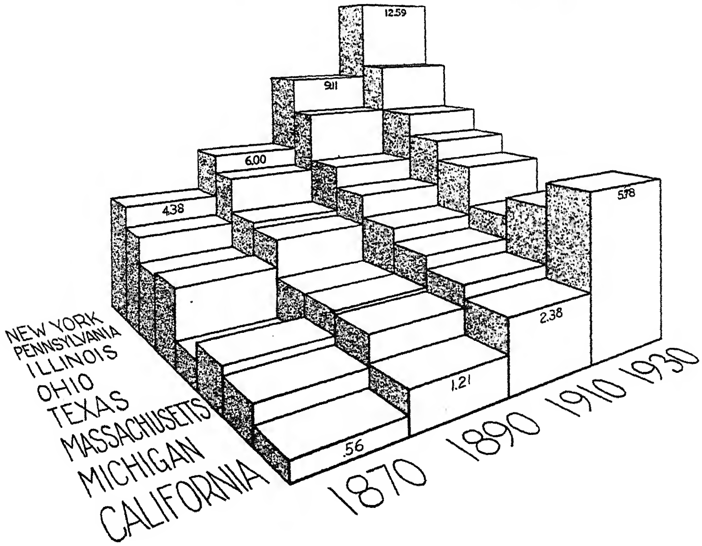
TABLE IV W

GROWTH IN POPULATION OF THE EIGHT STATES OF THE U.S. HAVING THE LARGEST POPULATION IN 1930

	POPULATION IN MILLIONS AT GIVEN DATES				
	1870	1890	1910	1930	SUMS
California	.56	1.21	2.38	5.78	9.93
Illinois	2.54	3.83	5.64	7.63	19.64
Massachusetts	1.46	2.24	3.37	4.25	11.32
Michigan	1.18	2.09	2.81	4.84	10.92
New York	4.38	6.00	9.11	12.59	32.08
Ohio	2.67	3.67	4.77	6.65	17.76
Pennsylvania	3.52	5.26	7.67	9.63	26.08
Texas	.82	2.24	3.90	5.82	12.78

CHART IV XVIII

GROWTH OF POPULATION OF EIGHT LARGEST STATES



blocks, is an arrangement based upon the "sums" column. We will therefore plan to have the tiers of blocks from rear to front, give the population data for the states in the order: New York,

Pennsylvania, Illinois, Ohio, Texas, Massachusetts, Michigan, California. Since there are eight states and only four dates, it will make for ease of lettering and reading if the axes are arranged as shown in Chart IV XVIII rather than in the reverse manner. A good general plan is to build up by means of sugar cubes, dominoes, or other blocks, an approximately correct structure, and then survey it from all angles until the vantage point is found yielding the clearest picture.

The actual steps showing the necessary construction lines involved in the construction of the rearmost block are shown in Chart IV XIX. In order they are as follows:

Draw horizontal line AB.

At some point, O, on it erect perpendicular CO.

At some point, D, whose height depends upon the height of the vantage point, but is in general higher than the highest block, draw horizon line FE parallel to AB.

On this line choose some point, E, for the right vanishing point and draw OE.

Choose some point, F, for the left vanishing point and draw OF. Angle EOF depends upon the distance of the vantage point above and in front of point O, the greater the height the smaller the angle and the greater the distance in front the larger the angle. This angle of necessity must be greater than 90° and less than 180°. An angle of 110° is frequently satisfactory, but a better estimate of it can be gotten by observing the angle in the domino structure when viewed from the desired point.

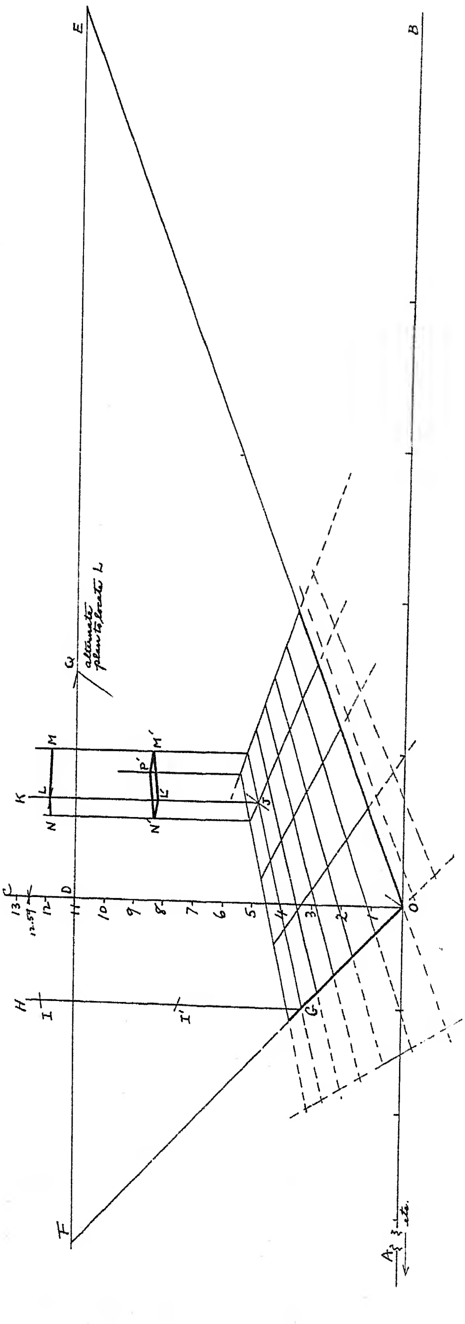
On OB lay off conveniently equally spaced intervals for the right horizontal scale.

On OA lay off conveniently equally spaced intervals for the left horizontal scale,

On OC lay off conveniently equally spaced intervals for the vertical scale, which must be

CHART IV XIX

CONSTRUCTION OF BLOCK DIAGRAM



such that the value for the tallest block falls below point D if it is desired to see the top of this block. The fact that the top of the tallest block cannot be seen in Chart IV XVIII is somewhat of a disadvantage.

Draw a grid representing the bases of the blocks by drawing lines from the division points on OB to F and from the division points on OA to E.

Draw the most remote block first, erecting perpendiculars from the four corners (three if the block extends above the horizon) of its base.

Erect HG, a perpendicular at G, the front corner of the rear tier of blocks.

On OC mark the proper height (12.59) for this most remote block.

Draw a line from this point (12.59) to F, cutting HG at I.

From I draw a line, ILME, to E, cutting JK at L, which is the upper front corner of the most remote block. LM is the upper right front edge and LN, a part of LF, is the upper left front edge. If the block extends above the horizon these two edges are all that can be seen, but if below the horizon the two rear edges are also to be drawn as shown by N′P′, a part of N′E, and M′P′, a part of M′F, had this rear block been 6.00 high instead of 12.59.

By similar procedure the remaining blocks in the rear tiers are constructed, erasing such of the lines already drawn as become covered by blocks in front.

Continue, always first constructing the most remote of the blocks remaining.

An alternative method for obtaining point L is to draw OJ and extend it to the horizon line FE, which it cuts at Q. Draw a line from point 12.59 on OC to Q. This line will cut JK at L. This alternative method is shorter to apply in most cases, but it breaks down if Q happens to be at the intersection of OC and FE.

If any block is completely hidden, it is necessary to start from the beginning again, using a different order in the rows, a higher perspective, or a different observation point.

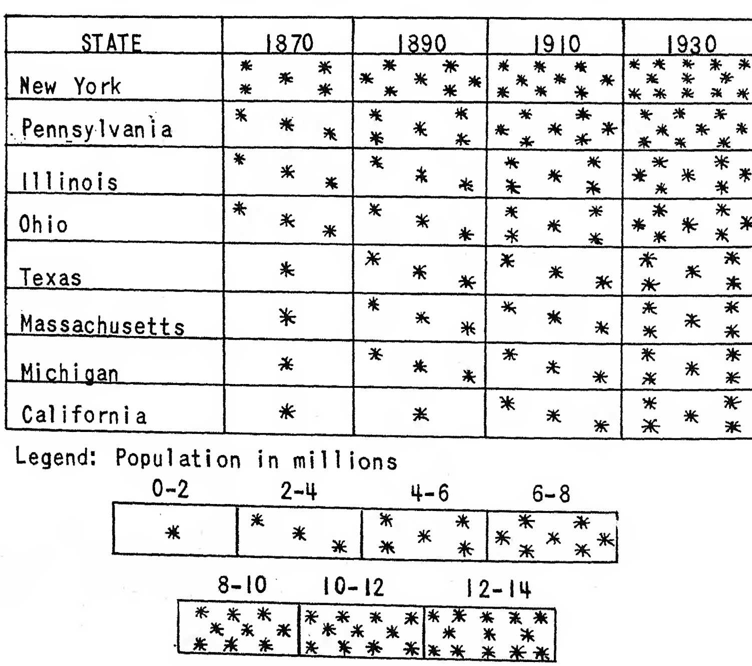
Drawing front edges slightly heavier, shading one face of the blocks, and recording heights on tops of blocks are frequently advantageous.

Label in perspective left and right horizontal dimensions, add title, and erase all construction lines.

The time scale is continuous and the blocks from left to right are appropriately in contact with each other, but as the states are discrete a diagram with small blank spaces between the blocks from front to rear would constitute a more accurate portrayal, but it would be much more difficult to draw than was Chart IV XVIII.

CHART IV XX

GROWTH OF POPULATION OF EIGHT LARGEST STATES



The cross-hatched plat: An alternative method of presentation of the growth in population data of Table IV W is by means of a plat with stipplings, shadings, or cross-hatchings to indicate the different frequencies, as shown in Chart IV XX.

The plat is more flexible than the block diagram method, but in general it requires a coarse frequency grouping and its features are not vividly depicted.

PROBLEMS

Problem 1. Plot temporal data of Table IV X.

TABLE IV X

CHINESE AND JAPANESE POPULATION IN THE U.S.: 1860 TO 1940*

CENSUS YEAR	CHINESE	JAPANESE
1940	77,504	126,947
1930	74,954	138,834
1920	61,639	111,010
1910	71,531	72,157
1900	89,863	24,326
1890	107,488	2,039
1880	105,465	148
1870	63,199	55
1860	34,933	—

*From the Sixteenth Census of the U.S., 1940

Problem 2. From 1919-20 data of Table IV Y compute the rate of mortality per year. Make necessary allowance in computations for the unequal age intervals reported. Plot male and female rate of mortality curves on same coordinate paper.

TABLE IV Y'''

LIFE TABLES FOR WHITE MALES AND FEMALES IN THE ORIGINAL REGISTRATION STATES*

EXACT AGE IN YEARS	1910#		1919-1920	
	MALES	FEMALES	MALES	FEMALES
0 months	100,000	100,000	100,000	100,000
1 month	95,156	96,213		
2 months	93,914	95,222		
4 months	92,039	93,632		
6 months	90,616	92,406		
8 months	89,453	91,394		
11 months	88,073	90,133		
1	87,574	89,774	90,757	92,639
2	85,201	87,455	89,050	91,070
4	83,449	85,812		
7	82,251	84,651	86,411	88,703
12	81,140	83,640	85,321	87,739
17	80,068	82,629	84,121	86,673
22	78,316	81,042	82,299	84,849
27	76,189	79,020	80,129	82,364
32	73,801	76,727	77,650	79,615
37	70,858	74,117	74,899	76,917
42	67,422	71,249	71,879	74,194
47	63,440	67,938	68,441	71,070
52	58,827	63,942	64,311	67,188
57	53,158	58,790	58,886	62,132
62	45,916	51,999	51,896	55,540
67	37,241	43,453	43,170	46,902
72	27,590	33,186	32,583	36,131
77	17,622	22,210	21,311	24,425
82	8,962	12,001	11,242	13,521
87	3,317	4,762	4,224	5,628
92	829	1,267	924	1,575
97	129	188		
102	10	14		

* The original registration states include New England states, New York, New Jersey, District of Columbia, Indiana, and Michigan.

\# Based on the estimated population July 1, 1910 (11,706,221), and on the reported deaths in 1909 (160,227), in 1910 (170,223), and in 1911 (165,918).

''' Reference: James W. Glover (1923) and Albertie Foudary (1923).

Problem 3. Examine Table IV Z. What seems to be the most striking phenomena revealed? Make graph to show this accurately.

TABLE IV Z

SCHOOL ATTENDANCE IN GEORGIA AND WASHINGTON BY AGE: 1940*

Age	Georgia		Washington	
	TOTAL NUMBER	ATTENDING SCHOOL	TOTAL NUMBER	ATTENDING SCHOOL
7 to 13 years	449,562	413,299	171,555	166,945
14 and 15 years	128,916	101,211	54,151	51,675
16 to 20 years	330,280	98,567	149,134	79,056
21 to 24 years	239,448	6,938	118,905	9,328

* Sixteenth Census of the U.S., 1940: POPULATION, Second Series, Characteristics of the Population, United States Summary, p. 76.

Problem 4. Choose the most appropriate interval possible and plot data of Table IV AA as a frequency polygon. Suggest an improvement in the original data which would have made still more comparable the performances of short and tall men.

Problem 5. Plot the data of Table IV AB as a percentile graph. We will first make certain nearly arbitrary assumptions: *(a)* That "under 1000" covers the range "400 to 1000"; *(b)* that "villages not incorporated" cover the range "30 to 400"; and *(c)* that "rural farms" cover the range "1 to 30." A community of size less than 1 is impossible so a zero percentile of .5 (the boundary between 0 and 1) may be plotted. The 100 percentile should not be plotted, since the maximum possible size of community is not determinable from this sample. Because of the great range, 1 to 6,930,446, use a logarithmic scale of sizes of communities (ordinate). The abscissa may

TABLE IV AA

STANDING HIGH KICK

Distances Measured to the Nearest Inch
Above or Below Standing Height (s.h.)

Score in Inches Above s.h.	Frequency
19	1
18	1
17	1
16	2
15	3
14	2
13	1
12	7
11	3
10	9
9	9
8	7
7	12
6	19
5	22
4	18
3	25
2	16
1	10
0	12
Below s.h.	
1	7
2	7
3	1
4	2
5	1
6	2
7	2
8	0
9	1
	$N = 203$

* Frederick Warren Cozens, THE MEASUREMENT OF GENERAL ATHLETIC ABILITY IN COLLEGE MEN, Univ. of Oregon Publication, Physical Education Series, Vol. 1, No. 3, April, 1929, page 188.

TABLE IV AB

URBAN-RURAL DISTRIBUTION OF POPULATION OF U. S. IN 1930*

SIZE OF COMMUNITIES	NUMBER	POPULATION
Cities		
6,930,446	1	6,930,446
1,000,000 - 4,000,000	4	8,134,109
400,000 - 1,000,000	13	8,067,471
200,000 - 400,000	23	6,508,600
100,000 - 200,000	52	6,685,110
25,000 - 100,000	283	12,917,141
10,000 - 25,000	606	9,097,200
5,000 - 10,000	851	5,897,156
2,500 - 5,000	1,332	4,717,590
Villages		
1,000 - 2,500	3,087	4,820,707
Under 1,000	10,346	4,362,746
Villages not incorporated		14,479,257
Rural Farms	6,288,648	30,157,513

* Abstract of the Fifteenth Census of the U.S., 1930 pp. 16, 17, 18, 19, 21.

be "Percentage of people residing in communities of smaller size than size indicated." Since 13 classes only are available if percentiles corresponding to the boundaries of these classes are computed and plotted, the result will be appreciably more accurate than if integral percentiles such as $P_{.01}$, $P_{.05}$, $P_{.10}$, etc., are computed and plotted.

Problem 6. Plot data of Table IV AC according to the five ways for drawing a graph mentioned in Section 3 (IV, p. 140) and compare results. Employ method *(c)* first in order to be uninfluenced by *(a)*, *(b)*, *(d)*, or *(e)*.

TABLE IV AC

HEAD LENGTHS OF 1306 NON-HABITUAL CRIMINALS*

HEAD LENGTH	FREQUENCY	HEAD LENGTH	FREQUENCY	HEAD LENGTH	FREQUENCY
210	1	196	66	182	28
209	1	195	66	181	17
208	0	194	83	180	13
207	6	193	79	179	12
206	3	192	102	178	7
205	8	191	96	177	5
204	13	190	85	176	3
203	14	189	83	175	3
202	24	188	68	174	0
201	20	187	55	173	2
200	30	186	57	172	1
199	35	185	53	171	1
198	43	184	43	170	0
197	56	183	24		
					1306

* Adapted from data given on page lxxv, TABLES FOR STATISTICIANS AND BIOMETRICIANS, edited by Karl Pearson, Vol. 1, 1914.

In employing method *(d)* we will smooth by replacing each original class frequency by the average of the 5 nearest class frequencies, e.g., 6, the frequency for class 207 is replaced by $3.6 \ (= \frac{1+0+6+3+8}{5})$ etc., for all classes from 168 to 213 inclusive.

Method *(e)* is very simple for these data because they are so nearly normal, but its execution may be postponed until later chapters, dealing with central tendencies, variability, and the normal distribution, have been studied.

Problem 7. Using data of Table IV AC, compute $P_{.05}$, $P_{.10}$, $P_{.25}$, $P_{.50}$, $P_{.75}$, $P_{.90}$, $P_{.95}$, and also the lowest and highest percentiles that are permissible.

Compute the standard errors of each of these percentiles.

Problem 8. Using data of Table IV W, select scales, left, right, and vertical, which are quite different from those of Chart IV XVIII, construct a block diagram. "Proofread" for accuracy by careful visual comparison of relative heights with those of Chart IV XVIII and by sighting down lines to see if they do converge as required.

Problem 9. The accompanying problem involves plotting contours and interpolating by means one degree more refined than is linear interpolation, and so as to preserve the frequencies in intervals rather than to preserve the frequencies at specific points,—class indexes. It is entirely feasible to present the data of Table IV AD in a block diagram, but they can also be well shown by frequency contour lines on a mat, the horizontal and vertical dimensions of which are age and grade. Lines for frequencies 100, 1000, 5000, 10,000, 15,000, 20,000, 25,000, 30,000, and 35,000 will suffice.

One may interpolate in the figures given by rows, and also again in the figures given by columns, to obtain grade and age points for each frequency, 100, 1000, etc. Connecting all the points for 100 gives a contour line for this frequency, and similarly for the other contour lines. Linear interpolation for points on the contour line for the frequency 100, though not altogether satisfactory, may suffice, but it

TABLE IV AD

AGE-GRADE DISTRIBUTION OF CALIFORNIA SCHOOL CHILDREN

(Compiled by Research Division of the University of California)

AGE AT LAST BIRTH-DAY	GRADE								
	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	
21.5	13	6	7	9	5	2	39	6	87
20.5	1	4	3	5	3	9	17	18	60
19.5	8	5	9	7	15	14	35	71	164
18.5	14	16	23	39	38	57	139	319	645
17.5	30	33	62	102	150	255	575	1340	2547
16.5	67	108	146	305	550	928	2176	4903	9183
15.5	123	168	317	686	1279	2621	5823	10791	21808
14.5	203	283	600	1417	2715	6151	13517	16531	41417
13.5	255	538	1113	2603	5959	11451	17564	11644	50927
12.5	700	1011	2373	5736	11685	18353	12402	2704	54964
11.5	889	2055	5271	11103	19509	12844	2708	251	54630
10.5	1912	4468	12217	21458	14377	3091	210	8	57741
9.5	4169	11288	24241	15875	2602	166	9		58350
8.5	11373	26688	18213	2562	136	4			58976
7.5	34609	19598	2084	106	2				56399
6.5	31206	1167	34						32407
UNDER 6.0	1203	15							1218
TOTAL	86775	67451	66713	62013	59025	55946	55014	48586	501521

hardly will for the other contour lines for the other contour lines for the age and grade groupings are coarse and the frequencies large. Since "under six "includes four- as well as five-year olds, neither the "under six" frequencies, nor any involving them in interpolation will be entirely satisfactory. The error in treating all "under six" cases as five-year-olds should not be great. Since no frequencies are given for kindergarten classes, no interpolated frequency values involving grade .5 (middle of the kindergarten year) will be satisfactory. Because of the lack

of kindergarten data neither linear nor parabolic interpolation as here described should involve grade .5. Consider the four sequential frequencies 0, 1203, 31206, and 34609. The five contour lines for frequencies of 5,000, 10,000, 15,000, 20,000, 25,000 will pass between the tabled values 1203 and 31206. Linearly interpolated values would be undesirably and unnecessarily crude. Parabolically interpolated values would yield values lying on a parabola fitted to three points, 0, 1203, and 31206, or 1203, 31206, and 34609. To avoid the issue as to which set of three to employ and to get at the same time somewhat more trustworthy results, we may fit a parabola to the points 1203 and 31206 and as near as possible, in the least-squares sense, to the neighboring points 0 and 34609. The equation of such a parabola is

$$f = \frac{1}{16}(-f_1 + 9f_2 + 9f_3 - f_4) + (f_3 - f_2)\xi + \frac{1}{4}(f_1 - f_2 - f_3 + f_4)\xi^2 \quad \ldots\ldots\ldots\ldots [4:04]$$

in which f_1, f_2, f_3, and f_4 are the four tabled frequencies in order and ξ is a deviation in years from 6.0, i.e., a point halfway between the class index values for the second and third tabled values. From this equation ξ may be determined for the needed contour levels, $f = 5000$, $f = 10{,}000$, $f = 15{,}000$, $f = 20{,}000$, $f = 25{,}000$, and $f = 30{,}000$. For example, for the 5000 level we have

$$5000 = \frac{1}{16}(0 + 10827 + 280854 - 34609) + (31206 - 1203)\xi + \frac{1}{4}(0 - 1203 - 31206 + 34609)\xi^2$$

yielding $\xi = -.371$ or -54.2, this latter value being extraneous as it is obvious that the root desired will lie between $-.5$ and $.5$. The value $\xi = -.371$ is equivalent to age 6.629, the age for grade one at which the frequency is 5000.

Though this method should give fair results, the following may logically be expected to give somewhat better values. We first note that the area under any stretch of curve [4:04] corresponds to the number of cases in this interval. Let us find by integral calculus this area from $\xi = -1.00$ to $\xi = .00$. The answer is 1249, which differs from the correct value by 46. Also, in the next interval the number of cases under the parabola is 31252, which is in error by 46. We may avoid these discrepancies by fitting a parabola having the observed areas for each of the two middle intervals and coming as close as possible, in the least-squares sense to having the requisite areas in the first and fourth intervals. The equation of this parabola differs only in the constant terms from [4:04]. It is

$$f = \frac{1}{12}(-f_1 + 7f_2 + 7f_3 - f_4) + (f_3 - f_2)\xi$$
$$+ \frac{1}{4}(f_1 - f_2 - f_3 + f_4)\xi^2 \ldots\ldots\ldots\ldots [4:05]$$

which is fully as simple to use as [4:04]. For the frequency level 5000 we have

$$5000 = \frac{1}{12}(0 + 8421 + 218442 - 34609) + (31206 - 1203)\xi$$
$$+ \frac{1}{4}(0 - 1203 - 31206 + 34609)\xi^2$$

yielding $\xi = -.370$ and $-.54.2$, this latter value being extraneous. The value $\xi = -.370$ is equivalent to age 6.630, which happens to be negli-

gibly different from 6.629 derived by the preceding method.

The determination, using [4:05] of all other contour points and the plotting of contour lines is left as an exercise. For checking purposes a few values for the contour level 5000 are given herewith: (age 7.5, grade 3.287), (age 8.5, grade 4.322), (grade 1.5, age 6.30 and also age 9.161), (age 6.5, grade 2.240). In the calculation of this last value we have the data

	Grade 1.5	Grade 2.5	Grade 3.5	Grade 4.5
Age 6.5	31206	1167	34	0

The contour point for level 5000 lies between grade 1.5 and grade 2.5, but as we have no recorded value for grade .5 we cannot use the method just illustrated. However, we can derive the equation of a parabola, reproducing exactly the areas in each of the first three intervals. If the origin is 2.00, i.e., a point midway between the first and second class indexes, and the frequencies in these three classes called f_2, f_3, and f_4, *the* equation is

$$f = \frac{1}{6}(2f_2 + 5f_3 - f_4) + (f_3 - f_2)\xi + \frac{1}{2}(f_2 - 2f_3 + f_4)\xi^2 \qquad \text{[4:06]}$$

For the data in hand this is

$$f = 11368\tfrac{5}{6} - 30039\xi + 14453\xi^2$$

which, when $f = 5000$, gives $\xi = .240$, corresponding to grade 2.240.

It is interesting to note in passing that, in a situation wherein f_1, f_2, f_3, and f_4 exist, if

f as given by [4:06] and f as given by a similar equation involving f_1, f_2, and f_3 are averaged, the resulting expression for f is [4:05].

Problem 10: Vocabulary Review

bar diagram
basal year
block diagram
categorical
center of population
circle chart
class frequency
class index
class interval
class limits
contingency table
contour lines
cumulative frequencies
cycle
frequency polygon
geographic
golden mean
grid
growth curve
histogram
horizon
inflexion, point of
least squares
linear interpolation
logarithmic chart
median
minor mode
mode
normal population
ogive
parabolic interpolation
percentile
percentile curve
perspective
pictogram
pie chart
plot
price index
pseudo growth curve
quadratic mean
qualitative
quantitative
relative time chart
sample
scatter diagram
segmented bar
semi-logarithmic chart
skewed curve
spatial
standard deviation
standard error
temporal
time chart
trend
vanishing point
zero point, natural
σ (sigma)
Σ (capital sigma)
ξ (xi)
χ^2 (chi square)
P_p
p
q
V_p
f_p
i_p
F_p
i_p'

CHAPTER V

THE STABLE FEATURES OF PHENOMENA

SECTION 1. THE QUEST FOR CERTAINTY

The yogi seeking peace in Brahma, the philosophical idealist finding reality in his own world of concepts, and the mathematician delving ever deeper into the abstractions of pure mathematics are united in abjuring the turmoil of human passions and the strife and struggle of competitive life and each is seeking in his separate way to find an anchor, a Rock-of-Ages, an everlasting law of being. The stabilities of life that each finds, or thinks he finds, are self-satisfying and eternal in the mind of the conceiver. Without questioning the existence and reality of stabilities of this sort we can note that their source is strictly and individually psychological, depending from the logical, instinctive, and habitual mental processes of the individual. We may also note that the applicability of universal and eternal principles and beliefs to concrete issues is never unquestioned, but in each instance subject to a judgment of pertinence. The pertinence has to be established by observation or experience and not by pure

deduction. The concepts of stability referred to, but not their fields of practical import, have a nonstatistical basis. Let us freely grant that this field of life does exist, and that it has an importance as judged by different individuals from supreme to trivial, and that it lies beyond the pale of experimental statistics.

The need and the search for anchors in life is not limited to philosophical idealists. They are, perchance, in less need for them than other folk, including avowed pragmatists and empiricists. The stabilities that these latter seek differ in three important respects from those sought by the former,—they are relative, not absolute, they are observationally determined, and they serve a finite purpose in an observably limited realm. This is the field served by applied statistics.

The concept of relatively ultimate durability or stability is valuable. Suppose it is reported that "Herman Rosenwald is a ten-year-old Jewish boy having an intelligence quotient of 130." The age item may be reacted to in a fixed manner for such a period of time as may be quite sufficient for many immediate issues. That his intelligence quotient is 130 is more stable temporally than was his 10-year-oldness. But this concept, consequent to some fallible measurement device applied at some moment in time, must be thought of as only relatively stable and only so with reference to certain none too clearly defined fields of functioning. That he is Jewish has greater temporal stability, but is of variable utility if used as an interpretative aid in delimiting mental and emotional conditions and activity. Though each of these items has relative, not ultimate, stability, nevertheless each is undoubtedly serviceable within limits, which are discoverable with greater or less precision.

Statistical study is primarily a search for

features of relative stability in connections not known from earlier study to possess them. This search should not be indiscriminate.

SECTION 2. BIOLOGICAL AND SOCIAL STABILITY

That there is form to life has been proclaimed by philosophers both early and late, and that it is rooted in biological, psychological, and social laws of survival and accommodation has been clearly pointed out by Charles Darwin, Lester Frand Ward, and others. The revelation of such form is the prime purpose of statistical study. Phenomena are conditioned in certain inherent ways which, though not immediately obvious, are compelling in their power to limit the range of values, and the relative frequencies of the different possible values, which can be taken.

If a cow is tied to a stake by a stout elastic tether, she is constricted in her grazing just as really as if tied with a Manila rope. Generally speaking, the restrictions upon social and biological phenomena are elastic, not rigid, and the strengths, or lengths, or existence, of these elastic bonds must be inferred by a study of tendencies, not of rigid limits. Can we infer the restricting effect of the elastic tether by observations as to the thoroughness with which the turf has been grazed? We can, but, primarily, not by noting the most extreme tufts plucked. Such a tuft might have had so succulent an appeal that the cow charged for and attained it at a distance quite beyond her usual grazing radius. By a careful study of the stubble we can obtain evidence not only as to the center of the grazing area, but also as to the elasticity or variability of the restrictive influence. If cow, tether, and stake are removed, and stake-hole obliterated, we can approximately reconstruct

the situation from the evidence yielded by the stubble, making close estimates of the location of the stake and of the variably restrictive influence of the tether.

Many phenomenological situations are characterized by a central tendency with elastic bounds. An oyster growing a thicker and thicker shell is more and more adequately protected from enemies but also more and more encumbered in movement. There is thus an optimum shell thickness—that is, a central tendency-coupled with decreasingly permissible variability from it—that is, a typical pattern of variability. The interest rate on a mortgage of certain amount, with 100 shares of stock A as collateral, would not be a fixed rate if there were many independent placements of such a mortgage. The lower the rate the fewer lenders, and the higher the rate the fewer borrowers, with the result that there would be a central tendency and a typical pattern of variability from it. Children of a given age are distributed throughout a number of school grades. They have a central tendency of scholastic ability and there exists a central tendency in school grade attended. As their abilities are farther and farther below the average there is greater and greater pressure by school authorities to lower the school grade attended, and as they are more and more able there is greater and greater individual effort to secure extra promotions so as to profit by higher grade offerings. Again a central tendency and a typical pattern of variability. Innumerable illustrations can be given.

There is no logical reason for believing that successive increases of .1 millimeter in thickness of oyster-shell above the central-tendency value will have a protective influence just equivalent in survival terms to the mobility benefit of successive decreases of .1 millimeter

below the central-tendency value. The detrimental effect of increased shell weight is conditioned by certain environmental conditions (perhaps getting food) which determine the variability pattern upon the above-average-shell-thickness side of the distribution, while the detrimental effect of decreased shell weight is consequent to other environmental conditions (perhaps escaping enemies), and this may well lead to a different variability pattern upon the below-average side of the distribution. The net result is a nonsymmetrical or skewed distribution. The disinclination of lenders to lend at low interest rates is a social phenomenon of a different sort from that of the disinclination of borrowers to borrow at high interest rates, so again we should expect the patterns of variability below and above the average interest rate to differ. Certainly the school pressure to demote the dull is of a different order from that of the individual urge to profit by extra promotions if bright.

We conclude that the skewed distribution is to be expected in physical, biological, psychological, and social life. We may also conclude that a knowledge of central tendency, of variability, and of skewness,—these last two together being a convenient summary of knowledge separately of variability below the average and above it,—is knowledge of three characteristic and quite independent aspects of phenomena which yield distributions. An equivalent statement, is that these aspects are stable features of phenomena of a certain type, recurring with slight modification from sample to sample.

Compare the enlightenment which the statement, "12½-year-old Bessie is in the low seventh grade," carries to the untutored and to the statistically informed auditor. The reaction of the first may be "What of it?" and of the second, that "Bessie

is bright and ambitious; she and her parents have progressively through a number of years fought the lock-step regimen of the school; and either entered school at an early age and was bright enough to profit by it or has been doubly promoted twice, for she is two half-grades above the average, a status attained by less than twenty per cent of her age-group; and her scholastic achievement is about equal to that of average low eighth-graders." The last of these items, and in part the first, depends upon knowledge of a statistic other than the three mentioned,—namely, that children who get extra promotions tend to get them to a grade location half as far above the average grade for their age as their ability warrants. The statistically informed has had knowledge of four stable features of age-grade-achievement phenomena that the untutored lacked. Furthermore, there are no substitutes for this knowledge, for, lacking the information inherent in the four statistical items mentioned, one would find it impossible to reach the same conclusion with equal certainty. It of course must, in connection with a deduction from these four items, as in the case of a deduction from any other information, be recognized that there is an element of uncertainty in a conclusion drawn, e.g., that from the facts given Bessie has the average scholastic ability of low eighth-graders has a probable error of approximately one-half a year.

Statistics should enable one to see the individual event in a setting and thus to know it as only he can who knows the background. The notion that statistical study deadens sensibility to individuality is utterly false. A person is more truly individualized when a given set of items about him are seen in the perspective of the population to which he belongs than when viewed as a unique event. To claim complete individu-

alization when viewed as a unique event asserts the ignorance of the observer rather than the individuality of the subject.

The balance that biologists repeatedly find to exist in the forms of life under a certain environment represents a type of stability discovered, described, and attested to with a known certainty by statistical study. The concept here is not stability within, or of, a single distribution, but stability of frequencies of occurrence of items of different sorts. The simplest case exists when two forms of life interact existentially. Ecology and parasitology are sciences concerned with stable relationships between living things. If a weevil, host to some parasite, exists in large numbers, the living conditions of the parasites are more favorable and they multiply to the point of excessive killing of their hosts, with the result that they must die in large numbers from duress of living conditions, with the result that the hosts now multiply, and so it continues until some natural balance is reached. Clearly, in this situation, the relative frequencies of host and parasite are crucial items of information; especially informative are these frequencies under stable life conditions.

The first concern of marine biology is perhaps the balance in the frequencies of the various forms of life in the homogeneous environment that the ocean provides. This balance can be highly trusted unless some new factor, such as man's predatory activities, changes things. In a certain social structure there can exist to mutual benefit so many milkmen to so many grocers, to so many barbers, to so many shoemakers, etc., and if some number should happen to be excessive it will tend to right itself. This same type of phenomena is precisely present in sundry chemical reactions.

Throughout the entire field of physical, biological, and social life we find that the frequencies, or proportions, in classes tend, under stable conditions, to fixed values. The trustworthy discovery of these proportions in classes is a fertile field of statistics. It is a phenomenological stability other than that discussed under form of distributions. However fixed these proportions are under stable conditions, they show typical growth or cyclical features under changing conditions. The proportion of people engaged in airplane manufacture has hardly attained such a value that we can look upon it as approximately fixed from year to year. For comprehension we require fixed reference points. If the proportion engaged in airplane manufacture does not provide one we look further. We might investigate if the rate of change of this proportion is constant, or geometric, or definable in some demonstrably precise terms. We must find some anchorage or expect that none but hazardous appraisals can be made. The logical need for awareness of the stabilities in phenomena is the need that statistics serve.

In studying proportions in classes and seeking evidence of stability therein it would seem necessary that the individuals in the classes be in some sense, or to some degree, interdependent. A study of the relative frequencies of cotton boll weevils and their parasites would be profitable, but a joint study of cotton boll weevils and citrus tree scales would not. Just as there are conflicting and interacting influences at work in determining the form of distribution, so there are in the matter of frequencies in classes. We are interested in those situations where these frequencies do in some sense directly interact or are affected by common causes.

In choosing cases for statistical study it is safe to say that the appropriate me*mbers should*

always be competitive or interacting individuals. A certain study of the intelligence of hoboes and college graduates can be cited. The relationships reported, of course, had no guidance or classificatory importance, and time has shown them to have been very misleading to readers unaware of the triviality of relationships between noncompetitive groups.

We will attempt to classify the sorts of stability possible in phenomena, admitting the likelihood of still other sorts not yet discovered or conceived.

SECTION 3. TEMPORAL DATA

Observations over a period of time may show (1) continuity of value from one moment to the next, (2) a general trend, (3) periodic fluctuations, (4) nearly constant proportionate relationships to some other variable, or variables, (5) evidences of limits below which, or beyond which, values cannot occur, (6) simplicities in any of the preceding five when some transformation of the time variable is made. Presumably, knowledge of the existence of a temporal series is direct and initial. That is all that is necessary to suggest the examination of the data along one or more of the six lines mentioned, with the hope of discovering its stable features. As soon as a type of stability is discovered, a foundation is laid for a study of causes and meanings.

SECTION 4. GEOGRAPHICAL DATA

Observations which are intrinsically connected with position in two-dimensional space may show (1) continuity of value from point to neighboring point, (2) discontinuity from point to neighboring point, (3) geographic patterns of frequency, (4) geographic patterns of categories, (5) direct or (6) inverse, frequency relationship from point

to point, (7) patterns in combination types of frequencies, (8) simplicities in any of the preceding seven under transformations of space. Knowledge of the existence of geographical data is direct and initial and immediately suggests search for stabilities of the eight sorts listed and for their causes and meanings.

Temporal and geographical: Many situations proclaim themselves as both temporally and geographically conditioned, as for example do farm crops, and then stabilities involving all combinations of the sorts in the two types listed are within the realm of possibility. The problem of discovery of stable features is greatly complecated, but the general modes of search are indicated by the 6 X 8 combinations.

SECTION 5. QUANTITATIVE AND QUALITATIVE DATA

Much data may be judged to be relatively independent of the time and place of collection in the sense that the issues involved are not dependent upon them, as, for example, would be a candidate's record upon a Civil Service examination. With data of this sort we look for stable features in the form of single distributions and in the relationships of two or more paired distributions.

Qualitative series: The fundamental types of stability are (1) proportions in classes as successive samplings are taken, (2) relationships between these proportions, (3) existence for these proportions of upper and lower limits other than one and zero, (4) relationships of super- and subordination between classes,—such relationships may be invariable or tendential.

A combination of stabilities of types (2) and (4) is revealed in many genetic chromosomal studies and similar conditions are to be expected in psychological vocational studies.

The statistics of sampling, association, and contingency are well developed and provide keen-edged tools for analysis of qualitative data.

Quantitative series: For quantitative series the fundamental types of stability reside in the parameters, or constants, which are definitive of a distribution or of distributions in their relationships. These are measures of (1) central tendency, (2) variability, (3) skewness, (4) kurtosis, (5) multimodality, (6) limits, (7) excluded values, (8) stability of each as affected by principle of sampling employed, (9) stabilities revealed as transformations of the scale of measurement are made, (10) relationships in all these respects of two or more paired series.

Generally items (1), (2), (3), (4), and (5) are of very unequal excellence in the information they give about a parent population. Many distributions are such that a sample of N cases will yield trustworthy information about, say, (1) and (2) and not about (3), (4), or (5). The higher moments μ_3, μ_4, μ_5, etc., have been proposed as descriptive statistics in connection with extremely skewed and leptokurtic series when not as justified upon the ground of stability as other neglected statistics. No matter how neat a statistic may be algebraically, its real merit is dependent upon its stability, the precision with which it depicts a feature of the population.

Amateurs err in coining and using an odd statistic because of some simplicity, real or fancied, in meaning, irrespective of and in ignorance of its instability. In illustration may be mentioned "the percentage of those getting A in mathematics who also got A in law" as a measure of correlation.

It must be admitted that professionals have at infrequent times erred in using a statistic

because of amenability to algebraic treatment irrespective of, though scarcely in ignorange of, its instability. In illustration may be mentioned the use of higher moments μ_3, μ_4, and even μ_8, though other measures such as percentiles and proportions in classes could reveal otherwise undisclosed though trustworthy features of the population.

Of two alternative statistics, each of sufficient reliability for samples of the size in question, we may well choose the one of greatest algebraic simplicity but never employ an unreliable statistic no matter its algebraic simplicity.

Quantitative and qualitative: Many situations combine the separate issues of the quantitative and qualitative series. One combination, well-nigh universal, is found when quantitative differences inhere in the very trait that has led to categorization, if one class of fruit fly is labeled "stubby wing" it will nevertheless be found that there are different degrees of stubbiness in those placed in this class; even in that class of people labeled "male" it is found that a variability in maleness exists. The same issue arises in quantum-theory physics. *The standard statistical approach to this problem is first to abstract all the information possible by treating the data as qualitative, and then to abstract still more information by superimposing a quantitative treatment on this.* Where such added quantitative treatment is not immediately feasible, the student should hold the strong conviction that he has but partially,—we may even say but superficially,—plumbed the issues. In the field of psychology the quantification of phenomena earlier thought of as qualitative has been one of the most fruitful means of advance. We may expect it to continue to be so both in

psychology and in other fields of physical and social science.

Another common combination of the two is found when quantitative and qualitative series enter into the single problem, as for example is the case where a person's intelligence (mainly quantitative as commonly measured) and his sex (qualitative as commonly measured) affect his fitness for some task. Though such bastard devices as tetrachoric and biserial correlation have been developed to cope with this situation, the student cannot help but feel that these measures hold but a part of the information necessary for a true* solution of the problem.

The qualification of seemingly quantitative phenomena is also a frequently illuminating process. The need for it would seem to exist where the so-called quantitative measure is largely subjective or hastily conceived, as for example is the case with measures of "general intelligence."

SECTION 6. COMPLEX DATA

Only the simpler types of complexity have been mentioned: temporal-geographical and quantitative-qualitative. However, very real and important problems involving all combinations of the various series exist. It is, in fact, quite arguable that every real problem does involve them all, and never one, two, or three only. There are two standard approaches to the handling of such problems. The generally preferred and more convincing one is so to set up experimental situations that variability in phenomena due to all but one of the types is negligibly small, so that the statistics of a single series may be employed. The other method employs the analy-

* Ascertainable "truth" being relative, the word is used to indicate higher rather than a lower stage, and not an absolute attainment.

sis of variance techniques and allocates to each type the variability attributable to it. This is a beautiful and powerful technique, but its merit is sufficiently restricted by its necessary complexity, when the data are complex, to warrant a serious search for experimental set-ups involving fewer issues.

SECTION 7. EMPLOYABLE INSTRUMENTS IN THE QUEST FOR CERTAINTY

Correlative with each type of stability is a statistic (or are statistics, in the case alternative methods are available) which meets the issue. These become apparent with statistical expertness. We give a few examples herewith: Temporal phenomena as listed in Section 3 numbers (1) and (2) are served by time charts, (3) by periodogram analysis and periodic trigonometric functions, (4) by ratios and index numbers, and (6) by growth curves and other transformations. Geographic phenomenon (1) is served by contour maps, (2) by spot maps, (3) and (4) by superimposed maps. Qualitative phenomenon (1) is served by proportions in classes and their standard errors (standard errors are important in all connections, but perhaps most strikingly so here), (2) by correlations between class frequencies in excess of that due to chance, (3) by association methods. Quantitative phenomenon (8) is served by Lexis' ratios and Fisher's variance-ratio tests, (9) by normalizing data and other nonlinear transformations, (10) by sundry measures of simple and multiple correlation, (9) and (10) by multivariate analysis, including mental factor analysis.

This list is in no sense exhaustive, but just a start which the student will find it profitable to amplify as his knowledge of techniques expands.

CHAPTER VI

MEASURES OF VARIABILITY

SECTION 1. THE GENERAL CONCEPT OF VARIABILITY

It is axiomatic that if all measures of a series are and must be alike there is nothing to investigate or report. The primary phenomenon of a series is that there are differences between the measures. We will later see that a deduction from this is that the measures have a mean (see [6:03]). It seems to the writer that the concept of variability is primary and that of central tendency secondary, but, as will be shown, one cannot think of the one in any but the most casual manner without being compelled to think of the other. They are as wedded as are the facts that a right triangle has three sides such that the sum of the squares upon two is equal to that upon the third, and it has three angles such that the sum of two is equal to the third. We now turn our attention to certain necessary relationships between the measures in a series.

To assist a student who has difficulty in dealing with algebraic symbols, a numerical illustration of quantities represented in Tables VI A, VI B, and VI C immediately accompanies each of

these tables. This numerical illustration need not be noted unless desired. The illustrative series is composed of the following four measures, 8, 5, 2, and 1, which have a mean of 4, a variance of 7.5, and a variance of differences of 20.

We will, then, start with the most elementary concept of variability possible, namely, that two measures of a series X_a and X_b differ. We will call the fact that $(X_a - X_b) = d_{ab}$, a quantity not zero, the most elemental or primitive measure of variability. If we have a series $X_a, X_b, X_c, \ldots X_n$, there are many such measures as given in Table VI A herewith, in which the null measures of the type $(X_a - X_a)$ must not be counted, but are included because they lead to later simplifications. We can let X stand for any measure of the series $X_a, X_b, X_c \ldots X_n$, but when it is necessary to distinguish between two such measures we require a different notation and shall let X_i stand for any of the X's and X_j any measure not X_i and we shall let $X_j{}'$ represent any measure including X_i. We clearly desire, as a comprehensive measure of variability for the

TABLE VI A

ALL POSSIBLE DIFFERENCES FROM A SERIES OF N MEASURES			ILLUSTRATIVE NUMERICAL DATA			
$X_a - X_a$	$X_b - X_a$	$X_c - X_a \ldots X_n - X_a$	8-8	5-8	2-8	1-8
$X_a - X_b$	$X_b - X_b$	$X_c - X_b \ldots X_n - X_b$	8-5	5-5	2-5	1-5
$X_a - X_c$	$X_b - X_c$	$X_c - X_c \ldots X_n - X_c$	8-2	5-2	2-2	1-2
.	.	. ... .	8-1	5-1	2-1	1-1
.	.	. ... .				
.	.	. ... .				
$X_a - X_n$	$X_b - X_n$	$X_c - X_n \ldots X_n - X_n$				

entire series, some function of all possible differences $(X_i - X_j)$. There are N^2 differences in Table VI A, but N of them, being null, do not count. We may thus write down the following generalized measure of variability:

$$f = \frac{\sum^{n^2-n} |(X_i - X_j)|^m}{N^2 - N} \qquad \text{A generalized measure of variability} \quad [6:01]$$

For typographical reasons N when a superscript or subscript is in lower case, but otherwise it is written as a capital. The n^2-2 over the Σ symbol indicates that there are N^2-2 terms in the summation. If nothing is written over the Σ, the reader is to understand that there are N addends.

This function, f, is worthy of investigation for different values of m. There may be beauties in measures of variability as yet undiscovered and unreported, but the writer has found the function rather intractable for all values of m except $m=2$. For this case $|(X_i - X_j)|^2 = (X_i - X_j)^2$ and the function has very simple and remarkable properties. The function, in this case, is a variance, for, by definition *a variance is the mean square of quantities whose mean is zero. (A standard deviation is the square root of a variance.)* Since $(X_i - X_j) = -(X_j - X_i)$ and for every $(X_i - X_j)$ in Table VI A there is in some other position a $(X_j - X_i)$, it follows that the mean of the differences in Table VI A = 0 and accordingly the mean of their squares is a variance. We may write

$$Vd = \frac{1}{N_2 - N} \sum^{n^2} (X_i - X_{j'})^2$$

The expression *Vd* stands for the variance of the differences and not for *V* times *d*. The symbol V_d is commonly used and is appropriate, but variances enter into formulas so frequently that it is advantageous to have them represented by a symbol without subscript or superscript, thus $V(d)$ is typographically simpler than its identical equal V_d or σ_d^2. The summation term is exactly the sum of the squares of the differences given in Table VI A. For the first column we have the magnitudes given in Table VI B.

TABLE VI B

SQUARES OF DIFFERENCES IN FIRST COLUMN OF TABLE VI A	ILLUSTRATIVE NUMERICAL DATA
$X_a^2 - 2X_aX_a + X_a^2$	$64 - 2 \times 8 \times 8 + 64$
$X_a^2 - 2X_aX_b + X_b^2$	$64 - 2 \times 8 \times 5 + 25$
$X_a^2 - 2X_aX_c + X_c^2$	$64 - 2 \times 8 \times 2 + 4$
$\vdots \quad \vdots \quad \vdots$	$64 - 2 \times 8 \times 1 + 1$
$X_a^2 - 2X_aX_n + X_n^2$	

Summing these we obtain the first row of Table VI C. The remaining rows being the sums corresponding to the second, third, etc., columns of Table VI A.

TABLE VI C

THE DIFFERENCES OF TABLE VI A SQUARED	ILLUSTRATIVE NUMERICAL DATA
$NX_a^2 - 2X_a\Sigma X_i + \Sigma X_i^2$	$4 \times 64 - 2 \times 8 \times 16 + 94$
$NX_b^2 - 2X_b\Sigma X_i + \Sigma X_i^2$	$4 \times 25 - 2 \times 5 \times 16 + 94$
$NX_c^2 - 2X_c\Sigma X_i + \Sigma X_i^2$	$4 \times 4 - 2 \times 2 \times 16 + 94$
$\vdots \quad \vdots \quad \vdots$	$4 \times 1 - 2 \times 1 \times 16 + 94$
$NX_n^2 - 2X_n\Sigma X_i + \Sigma X_i^2$	

The sum of all these yields $N\ \Sigma X_i^2 - 2(X_i)^2 + N\ \Sigma X_i^2$, yielding finally

$$V_d = \frac{2[N\ \Sigma X_i^2 - (\Sigma X_i)^2]}{N^2 - N} \quad \text{Variance of differences between the measures in a series} \qquad [6:02]$$

As, by definition, the mean, M, of a series is their sum divided by their number, we may write $\Sigma X = NM$ and [6:02] may be written

$$V_d = \frac{2[\Sigma X^2 - N\ M^2]}{N-1} \quad \text{Variance of differences between the measures in a series} \qquad [6:03]$$

As there is no ambiguity the i as a subscript of X has been dropped.

Though no mention of the mean was present in the statement of the problem, we find it turning up as an essential statistic in [6:03].

Computation by formula [6:03] is simple, whereas it would be laborious by squaring the values of Table VI A.

We may express V_d in different notation. *The variance of the X's is by definition the sum of the squares of their differences from their means divided by their number.* This follows from the easily proven and never-to-be-forgotten fact that $\Sigma(X-M)=0$. Thus by definition

$$V = \frac{\Sigma(X-M)^2}{N} \quad \text{The variance of a distribution} \qquad [6:04]$$

As a simple exercise the reader may prove that $\Sigma(X-M)^2 = \Sigma X^2 - N\ M^2$ so that

$$V = \frac{\Sigma X^2 - N\ M^2}{N} \quad \text{The variance of a distribution} \qquad [6:05]$$

Also by definition

$$\sigma = \sqrt{V} \quad \text{The standard deviation of a distribution} \qquad [6:06]$$

We may now write [6:03] thus

$$V_d = 2V\frac{N}{N-1} \qquad \text{Variance of differences between the measures in a series} \qquad [6:07]$$

The most widely used measure of variability is V or its square root σ, but, because of the factor $N/(N-1)$ there is not a constant relationship between V_d and V, though it is nearly constant if N is large. V_d is the preferred measure (see [6:08]) because $\bar{\tilde{V}}_d = V_d$ but $\bar{\tilde{V}} \neq V$.

The structure of the symbol $\bar{\tilde{V}}$ should carefully be noted. V stands for variance. Had one the true variance, or that for the infinite population of which the N cases are a random sample, we would write it $\tilde{V}$. We never have $\tilde{V}$ except hypothetically, for we never deal with samples of infinite size. *A bar over a symbol has one of two meanings,—it indicates either a mean or an unbiased estimate.* Thus $\bar{\tilde{V}}$ is an estimate of the population variance. *The tilde to represent a population value and the bar to represent an unbiased estimate or a mean will be used throughout this text.*

Though the variance, V, obtained from the sample, is not an unbiased estimate of the population variance, the variance of the differences given by the sample, V_d, is an unbiased estimate of the variance of the differences that would be found in the infinite population.

$$\bar{\tilde{V}}_d = V_d \qquad \text{Unbiased estimate of the variance of the differences between the measures in the population} \qquad [6:08]$$

The unbiased estimate of the population variance is a function of the sample variance and the number of cases in the sample, thus,

$$\bar{\tilde{V}} = \frac{N}{N-1}V = \frac{1}{2}V_d \qquad \text{Unbiased estimate of population variance} \qquad [6:09]$$

Letting $\overline{V}$ and $\overline{V}_d$ be the means of the V's and of the V_d's found in an infinite number of similar samples, [6:09] becomes [6:10] showing the relationships between true measures of variability.

$$\tilde{V} = \frac{N}{N-1}\overline{V} = \tfrac{1}{2}\overline{V}_d = \tfrac{1}{2}\tilde{V}_d \quad . \; . \; . \; . \; . \; . \quad [6:10]$$

Recapitulating the development we note *(a)* we started with the elementary concept that the variability of a series of measures is a function of all the possible differences between them; *(b)* that the algebraically simplest function of this sort is the one yielding the variance of the differences between the measures; *(c)* that the computational procedure for obtaining this calls for a new statistic, the mean; *(d)* that V_d is $2N/(N-1)$ times the variance, V, of the sample, which is a universally employed statistic; *(e)* that V_d itself is an unbiased statistic requiring no modification depending upon the size of sample employed; *(f)* that one-half V_d, or its equal $NV/(N-1)$, is an unbiased estimate of the population variance.

SECTION 2. DEGREES OF FREEDOM

Because of the simplicity with which it follows from properties of V_d, the matter of degrees of freedom is discussed at this point. If we have three independent measures X_a, X_b, X_c, knowledge of one conveys no information as to the values of the others, singly or jointly, so we have three degrees of freedom. In Table VI A there are N^2 measures. How many of these are independent? Obviously we know the exact value of the null measures so they provide no degrees of freedom. Having X_a-X_b gives us no information as to X_a-X_c, X_a-X_d. . . X_a-X_n, so there are exactly $N-1$ degrees of freedom represented by the first column. The first measure in the second column, X_b-X_a, is the negative of X_a-X_b and is thus not independent. The third measure, X_b-X_c,

is derivable from measures in the first column, for it equals $(X_a - X_c)-(X_a - X_b)$ and is thus not independent. Nor are any of the remaining measures in this column. Clearly there are no measures independent of those in the first column to be found anywhere in the entire table so, though there are N^2 measures in the table, there are only N-1 degrees of freedom. Also we may observe that since there are N-1 degrees of freedom in V_d, and since

$$V = \sigma^2 = \frac{N-1}{2N} V_d$$

there are N-1 degrees of freedom in V and also in σ.

In the computation of V we had $V=[\Sigma(X-M)^2]/N$ and we noted that this is a biased statistic, but if we divide the summation by the number of degrees of freedom instead of by N, we have $\tilde{\bar{V}} = [\Sigma(X-M)^2]/(N-1)$ which is an unbiased statistic. *In general, summations are to be divided by the number of degrees of freedom, not by the number of cases, to obtain unbiased statistics.*

Is the mean an unbiased statistic? $M = (X_a + X_b + X_c \ldots X_n)/N$. Since the various X's are independent, there are just N degrees of freedom and M is unbiased. We may write

$$\tilde{\bar{M}} = M \qquad \text{An unbiased estimate of the population mean} \qquad [6:11]$$

Certain statistical manipulations deduct more degrees of freedom than do others. For example, if we take differences of differences we have N-2 degrees of freedom, though the number of such is N^4 when we include the null values, and it is N^4-2N^3+N when they are excluded.

The student of statistics will find that this matter is important in connection with small samples and in connection with sundry derived

statistics wherein algebraic restrictions have been imposed upon the data.

SECTION 3. THE VARIANCE AND THE STANDARD DEVIATION OF THE MEAN

We found that V_d and V are interchangeable in the sense that having one, and the number of cases in the sample, we can immediately obtain the other. Thus either may be taken as a measure of variability of the data. Just as variability of the data is its chief phenomenon, so variability of the mean (not its absolute value) is its prime characteristic. What matters the value of a mean if it is so quixotic that it cannot be trusted? We shall now obtain the variance of the difference between means and the variance of means. If K is the number of samples, we have K means and obviously (see [6:07])

$$V(M_i - M_j) = \frac{2K}{K-1} V_M \qquad [6:12]$$

and as the number of samples approaches infinity we may write

$$\tilde{V}(M_i - M_j) = 2\tilde{V}_M \qquad [6:13]$$

The error in any single mean, say M_1, is $M_1 - \tilde{M}$. We call this Δ_1 and write $\Delta_1 = M_1 - \tilde{M}$. Clearly $\overset{k}{\Sigma}\Delta_i/K$ tends toward zero, for each M_i is an unbiased estimate of $\tilde{M}$. By definition of a variance

$$V_M \equiv V_\Delta = \frac{\overset{k}{\Sigma}\Delta^2}{K}$$

If $\tilde{M}$ is subtracted from each X in Table VI A, the values of the differences are unchanged for $[(X_i - \tilde{M}) - (X_j - \tilde{M})] \equiv X_i - X_j$. We now define $\tilde{x}$, a

score as a deviation from the population mean, by

$$\tilde{x} = X - \tilde{M} \quad \text{A deviation-from-the-true-mean score} \qquad [6:14]$$

For the first sample of N cases we have

$$\frac{\Sigma \tilde{x}}{N} = M_1 - \tilde{M} = \Delta_1$$

and similarly for successive samples. Since $\tilde{x}_i - \tilde{x}_j \equiv X_i - X_j$ we may substitute in Table VI A and continue the development as before, obtaining the first equation of Table VI D as the comparable formula to [6:03].

The subscript 1 has been added to indicate that measures from the first sample only are involved. For the K samples we have the equations of Table VI D.

TABLE VI D

GIVING VARIANCES OF DIFFERENCES BETWEEN SCORES FOR K SUCCESSIVE SAMPLES OF N EACH

$$(N-1)\, V_{d_1} = 2\,\Sigma\,\tilde{x}_1^2 - 2N\,\Delta_1^2$$

$$(N-1)\, V_{d_2} = 2\,\Sigma\,\tilde{x}_2^2 - 2N\,\Delta_2^2$$

$$(N-1)\, V_{d_3} = 2\,\Sigma\,\tilde{x}_3^2 - 2N\,\Delta_3^2$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$(N-1)\, V_{d_k} = 2\,\Sigma\,\tilde{x}_k^2 - 2N\,\Delta_k^2$$

We shall now sum these equations, as K approaches infinity, first noting three simplifications (1) that each of the V_d's is an unbiased estimate of $\tilde{V}_d$ and so clearly their average approaches $\tilde{V}_d$,

(2) that $\Sigma\,\tilde{x}_1^2 + \Sigma\,\tilde{x}_2^2 + \ldots + \Sigma\,\tilde{x}_k^2 = \overset{nk}{\Sigma}\;\tilde{x}^2$ which

approaches $NK\tilde{V}$, and (3) that $\overset{k}{\Sigma}\Delta^2$ approaches $K\tilde{V}_m$. Thus we obtain

$$(N-1)K\tilde{V}d = 2NK\tilde{V} - 2NK\tilde{V}_m$$

Utilizing the relationships $\tilde{V}_d = 2\tilde{V}$ and $\tilde{V} = [N/(N-1)\overline{V}$, we obtain

$$V_m = \frac{\tilde{V}}{N} = \frac{\overline{V}}{N-1} \quad \text{True variance error of the mean} \ldots\ldots [6:15]$$

If in place of the unknown $\overline{V}$ we use an unbiased estimate of it, we have the accompanying highly serviceable formula, in which V_m is the usual and shorter notation for $\overline{\tilde{V}}_m$.

$$V_m = \frac{V}{N-1} \quad \text{Variance error of the mean} \ldots\ldots\ldots\ldots [6:16]$$

It would be more accurate to designate this as the "approximate variance error of the mean," but as every variance error, or standard error, formula based on the observed statistics is approximate, we omit the word "approximate," though the student should always be aware that a sample estimate has been employed in lieu of the unknown population statistic.

Noting [6:12], we may also write

$$V(M_i - M_j) = \frac{2V}{N-1} \quad \text{Variance error of difference between means from samples of N each from a common parent population.} \ldots\ldots [6:17]$$

As the square root of any variance error is a standard error, [6:16] yields

$$\sigma_m = \frac{\sigma}{\sqrt{N-1}} \quad \text{Standard error of the mean} \ldots\ldots\ldots\ldots [6:18]$$

By a slight extension of [6:17] we have for two means, based upon samples of different size and drawn from different parent populations:

$$V(M_1-M_2) = \frac{V_1}{N_1-1}+\frac{V_2}{N_2-1} \quad \text{Variance error of difference between independent means} \quad [6:19]$$

Knowledge of V_M or σ_M is necessary to an understanding of the mean. Those who interpret means without knowledge of their standard errors assume, albeit unconsciously, that σ_M is small, for otherwise the mean is untrustworthy. For every refined judgment and for every debatable issue this essential item must not be assumed, and it need not be, for [6:18] is available and simple to use.

The reader may frequently find formulas for the preceding standard and variance errors having N instead of $N-1$ in the denominators. This procedure is not recommended (for V is a better estimate of $\overline{V}$ than it is of $\widetilde{V}$) though no error of interpretation is likely to result for samples even as small as $N=12$. A standard or variance error is never needed to more than two significant figures, and a one-significant-figure answer usually suffices.

SECTION 4. SAMPLE AND POPULATION MOMENTS

Since the mean and variance are so mutually cooperative in revealing stable features of a distribution, let us consider a computational method yielding both. The standard method of computation provides the basis for getting higher moments as well. We define these as follows:

$$\frac{\Sigma(X-P)^k}{N} \quad \text{The k'th moment from the fixed point P} \quad [6:20]$$

P may be any point not dependent upon the values of X in the sample. A common case arises when $P = 0$. If, in this case, $k = 1$, then obviously [6:20] yields the mean. In other words, the

mean is the first moment from the zero point of the scale of measurement of the X scores. If we replace P by the variable point M (variable because the mean changes from sample to sample) we have the *k'th* moment from the mean, which is usually referred to as simple the *k'th* moment. If moments from some point P, other than M, are involved, they must always be designated moments from P.

$$'\mu_m = \frac{\Sigma(X-M)^k}{N} \text{ The k'th moment} \qquad [6:21]$$

We note that the first moment = 0; that the second moment = V, which may be designated μ_2; that the third moment, designated μ_3, equals zero in a symmetrical distribution, but not otherwise, and thus is a measure of asymmetry; that the fourth moment, μ_4, measures a variability characteristic unrelated to asymmetry and different from $V(= \mu_2)$ in that it is more subject to the influence of extreme measures than is V. This phenomenon, measured by the quotient μ_4/V^2 (see [7:03]) is called the kurtosis of a distribution and is high when the distribution has long tails and a pronounced mode. Otherwise expressed, if the parts of the curve intermediate between the mode and the tails are called the hips, *low, intermediate, and high hip heights correspond to leptokurtic, mesokurtic, and platykurtic distributions.*

Higher moments beyond the fourth will not here concern us, though we will state without proof that the variance error of any moment depends upon moments up to one twice as high as itself. We have already found that the variance error of the first moment, the mean, depends upon V, the second moment. The variance of V depends upon moments as high as μ_4, and similarly for higher moments. An unhappy feature of the higher moments is their instability, as is im-

mediately obvious from the fact that a single measure, if extreme, will greatly affect their values.

In moment notation the μ's stand for moments from the mean, thus $\mu_1 = 0$, a quantity that never needs to enter into a formula. However, the mean does enter in, and it has been customary to represent it in moment notation* either as $\tilde{\mu}_1$ or

* The notation here used is related to that of Fisher (1925 et seq. and 1928) and to that of Yule and Kendall (1937) as follows:

AS HERE USED	AS USED BY FISHER	AS USED BY YULE AND KENDALL
N	n	N
X	x	X
ξ		ξ
Arb. or		A
x	at times $(x-\bar{x})$, at other times X	x
$M = \bar{\tilde{M}}$	$m_1' = k_1$	M
$V = \sigma^2 = \mu_2$	m_2	$\mu_2 = \sigma^2$
$V = \frac{N}{N-1} V$	$k_2 = \frac{n}{n-1} m_2$	
μ_3	m_3	μ_3
μ_4	m_4	μ_4
$\tilde{M}$	$\mu_1' = \kappa_1$	
$\tilde{V}$	$\mu_2 = \kappa_2$	
$\tilde{\mu}_3$	$\mu_3 = \kappa_3$	
$\tilde{\mu}_4$	$\mu_4 = \kappa_4 + 3\kappa_2^2$	

It will be noticed that all theoretical, or population statistics are here designated by tildes and that Fisher represents them by Greek letters. Fisher's sample k statistics are unbiased estimates of the population κ statistics and are closely related to Theile's seminvariants.

$\mu_1{}'$ or $m_1{}'$. We shall avoid these and represent the mean by M.

Except the mean, none of the raw moments are unbiased statistics. The accompanying formulas show the extent of this bias for all moments, and combination of moments, up to the fourth degree. A vinculum over a statistic, or product of statistics, indicates a mean value, obtained by averaging for very many samples of N each.

$$\overline{M} = \tilde{M} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad [6:22]$$

$$\overline{M^2} = \tilde{M}^2 + \frac{\tilde{V}}{N} \quad \ldots\ldots\ldots\ldots\ldots \quad [6:23]$$

$$\overline{M^3} = \tilde{M}^3 + \frac{\tilde{\mu}_3}{N^2} + 3\frac{\tilde{M}\,\tilde{V}}{N} \quad \ldots\ldots\ldots\ldots \quad [6:24]$$

$$\overline{M^4} = \tilde{M}^4 + \frac{\tilde{\mu}_4 + 3(N-1)\tilde{V}^2}{N^3} + \frac{6\tilde{M}^2\tilde{V}}{N} + \frac{4\tilde{M}\,\tilde{\mu}_3}{N^2} \quad [6:25]$$

$$\overline{V} = \frac{N-1}{N}\tilde{V} \quad \ldots\ldots\ldots\ldots\ldots \text{ See } \quad [6:10]$$

$$\overline{V^2} = \tilde{V}^2 + \frac{(N-1)^2}{N^3}\tilde{\mu}_4 - \frac{3N^2-5N+3}{N^3}\tilde{V}^2 \quad [6:26]$$

$$\overline{\mu_3} = \frac{(N-1)(N-2)}{N^2}\tilde{\mu}_3 \quad \ldots\ldots\ldots\ldots [6:27]$$

$$\overline{\mu_4} = \frac{(N-1)(N^2-3N+3)}{N^3}\tilde{\mu}_4 + \frac{3(N-1)(2N-3)}{N^3}\tilde{V}^2 \quad [6:28]$$

$$\overline{MV} = \frac{(N-1)}{N}\tilde{M}\ \tilde{V} + \frac{N-1}{N^2}\tilde{\mu}_3 \quad \ldots\ldots\ldots\ldots [6:29]$$

$$\overline{M^2V} = \frac{N-1}{N}\tilde{M}^2\tilde{V} + \frac{N-1}{N^3}[2N\tilde{M}\ \tilde{\mu}_3 + (N-3)\tilde{V}^2] \cdots\cdots [6:30]$$

$$\overline{M\mu_3} = \frac{(N-1)(N-2)}{N^2}\tilde{M}\ \tilde{\mu}_3 + \frac{(N-1)(N-2)}{N^3}(\tilde{\mu}_4 - 3\tilde{V}^2) \quad [6:31]$$

It is to be noted that for a normal distribution, in which $\tilde{\mu}_3 = 0$ and $\tilde{\mu}_4 = 3\tilde{V}^2$, most of the preceding formulas simplify greatly. After the student has studied correlation he will find the last three formulas useful in obtaining the correlation between mean and variance, between mean-squared and variance, and between mean and third moment.

From formulas [6:22], [6:10], [6:27], [6:28] and [6:26] we compute the following unbiased estimates of the first four population moments:

$\tilde{\overline{M}} = M$ An unbiased estimate of the population mean, see [6:11]

$$\bar{\tilde{V}} = \frac{N}{N-1} V$$ Ab unbiased estimate of the population variance, see [6:09]

$$\bar{\tilde{\mu}}_3 = \frac{N^2}{(N-1)(N-2)} \mu_3$$ An unbiased estimate of the population third moment [6:32]

$$\bar{\tilde{\mu}}_4 = \frac{N}{(N-1)(N-2)(N-3)} [(N^2-2N+3)\mu_4 - 3(2N-3)V^2]$$

An unbiased estimate of the population fourth moment [6:33]

SECTION 5. CUMULANTS AND k-STATISTICS

For problems involving the combinations of statistics (especially if of the fourth or higher degree in X) from different samples, the student will find substantial economy in technique and thought by employing Fisher's (1928 and 1934) k-statistics. The first four k-statistics are given herewith, in which μ_3 and μ_4 are as here defined and not as defined by Fisher (see footnote Ch. VI, p. 212).

$$k_1 = M$$ Fisher's k_1 [6:34]

$$k_2 = \frac{N}{N-1} V$$ Fisher's k_2 [6:35]

$$k_3 = \frac{N^2}{(N-1)(N-2)} \mu_3$$ Fisher's k_3 [6:36]

$$k_4 = \frac{N^2}{(N-1)(N-2)(N-3)} (N+1)\mu_4$$ Fisher's k_4

$$-3(N-1)V^2$$ [6:37]

These k-statistics are unbiased estimates of Fisher's kappa, or population, statistics κ_1, κ_2, κ_3, and κ_4, called cumulants.

The relationships between the first four cumulants and population moments are :

$$\tilde{M} = \kappa_1 \tag{6:38}$$

$$\tilde{V} = \kappa_2 \tag{6:39}$$

$$\tilde{\mu}_3 = \kappa_3 \tag{6:40}$$

$$\tilde{\mu}_4 = \kappa_4 + 3\kappa_2^2 \tag{6:41}$$

Some of the merits of cumulants and k-statistics can be surmised from the following quotation from Fisher (1934, p. 22): "it is to [Laplace] we owe the principle that the distribution of a quantity compounded of independent parts shows a whole series of features — mean, variance, and other cumulants — which are simply the sums of like features of the distributions of the parts."

In an important sense all the moments beyond the first are measures of variability, but they measure quite different phenomena and we shall follow the custom of referring to the standard deviation, the variance, or some comparable measure, when speaking of variability, and employ the terms skewness and kurtosis for such special aspects of variability as are revealed by asymmetry and height at the hips.

SECTION 6. THE COMPUTATION OF THE MEAN AND OF MOMENTS

Let us now consider computational methods for getting the moments. Raw scores, designated X-scores, and deviation-from-mean scores (sometimes abridged to "deviation scores"), designated x-scores, and standard scores, or deviation-from-mean-in-terms-of-the-standard-deviation scores, designated z-scores ($z=x/\sigma$), have their respective merits. (Standard scores are designated

x-scores in normal probability tables, for z is there used to indicate the ordinate. See [8:23]. Final outcomes must be expressed in terms of scores because they are the observed measures. Algebraic derivations are frequently simplified by employing x or z scores. None of these well serve most computational needs, for the benefits of grouping and employing non-fractional values are not attained by them. We therefore employ a new variable, ξ, which is a score as a deviation from an arbitrary origin in terms of the grouping interval, i, employed. We have

$$\xi = \frac{X - \text{Arb. or.}}{i} \quad \text{Relation between } X \text{ and } \xi \text{ scores.} \qquad [6:42]$$

or $X = \text{Arb. or.} + i\,\xi$

If M_ξ is the mean of the ξ scores, we also have

$$M = \text{Arb. or.} + i\,M_\xi \quad \text{The mean computed via } \xi \text{ scores} \qquad [6:43]$$

$$V = i^2 V_\xi \text{ also } \sigma = i\,\sigma_\xi \quad \text{Variability computed via } \xi \text{ scores} \qquad [6:44]$$

We also have

$$x = i(\xi - M_\xi) \text{ or } \xi = \frac{x}{i} + M_\xi \quad \text{Relation between } x \text{ and } \xi \text{ scores} \qquad [6:45]$$

From [6:45] we easily derive

$$V = i^2 V_\xi = i^2\left(\frac{\Sigma\xi^2}{N} - M_\xi^2\right) \quad \text{The variance computed via } \xi \text{ scores} \qquad [6:46]$$

$$\mu_3 = i^3 \mu_{3\xi} = i^3\left(\frac{\Sigma\xi^3}{N} - 3M_\xi \frac{\Sigma\xi^2}{N} + 2M_\xi^3\right) \qquad [6:47]$$

$$\mu_4 = i^4 \mu_{4\xi} = i^4\left(\frac{\Sigma\xi^4}{N} - 4M_\xi \frac{\Sigma\xi^3}{N} + 6M_\xi^2 \frac{\Sigma\xi^2}{N} - 3M_\xi^4\right) \qquad [6:48]$$

An important property of the arithmetic mean readily follows from [6:46] by considering the case in which the grouping interval equals one. Then

$$V = \frac{\Sigma x^2}{N} = \frac{\Sigma \xi^2}{N} - M_\xi{}^2 \text{ or } \Sigma x^2 = \Sigma \xi^2 - NM_\xi{}^2$$

immediately proving that $\Sigma x^2 < \Sigma \xi^2$ whenever M_ξ, the distance from arbitrary origin to mean, does not equal zero. That is, *the sum of squared deviations is a minimum when the point from which the deviations are taken is the mean.*

We now have two unique properties of the mean, $\Sigma x = 0$ and Σx^2 = a minimum, which are the cause of its algebraic simplicity and its general superiority as a measure of central tendency. Formulas [6:43], [6:46], [6:47], and [6:48] are desirable work formulas in that the arithmetic may be made simple by an appropriate choice of the arbitrary origin and the grouping interval. If the grouping is coarse there is a systematic error in the even moments (but not in the odd moments) due to grouping which, for certain forms of distribution may be allowed for by applying Sheppard's corrections as given in formulas [6:49], [6:50], wherein ${}_sV$ and ${}_s\mu_4$ are the moments after applying Sheppard's corrections.

$${}_sV = V - \frac{i^2}{12} \qquad [6:49]$$

Sheppard's corrections to the second and fourth moments

$${}_s\mu_4 = \mu_4 - \frac{i^2 V}{2} + \frac{7i^4}{240} \qquad [6:50]$$

Because of lack of universal applicability and because they would interfere with tests of significance, it is very desirable to employ such a grouping that the corrections are negligible and may therefore be omitted. Sheppard's correction

in σ amounts to one per cent when $i = .49\ \sigma$ and the correction in μ_4 is one per cent when i is approximately $.25\ \sigma$. As, for most of the purposes of social statistics, the standard deviation is the crucial measure of variability and a one per-cent error in this measure is immaterial, we will *in general endeavor to group for computational purposes so that i is not greater than .5 σ, which rule is substantially equivalent to one calling for 12 or more classes.* The reader will appreciate that in situations calling for more than common precision and in situations wherein the argument depends from higher moments than V, either a finer grouping than this should be employed or, in cases not involving tests of significance, Sheppard's corrections employed.

The computation of the first four moments for the temperature data of Table IV J, $i = 3°$ column, is shown herewith.

TABLE VI E

THE COMPUTATION OF M, V, μ_3, AND μ_4

X	f	ξ	$f\xi$	$f\xi^2$	$f\xi^3$	$f\xi^4$
99	2	5	10	50	250	1250
96	2	4	8	32	128	512
93	0	3				
90	1	2	2	4	8	16
87	6	1	6	6	6	6
84	13	0	(26)		(392)	
81	23	-1	-23	23	-23	23
78	5	-2	-10	20	-40	80
75	6	-3	-18	54	-162	486
72	1	-4	- 4	16	-64	256
69	1	-5	- 5	25	-125	625
66	2	-6	-12	72	-432	2592
	62		(-72)	302	(-846)	5846
			-46		-454	
	N		$\Sigma\xi$	$\Sigma\xi^2$	$\Sigma\xi^3$	$\Sigma\xi^4$

$$M_\xi = \frac{\Sigma\xi}{N} = -.7419$$

$$V\xi = \frac{\Sigma\xi^2}{N} - M_\xi^2 = 4.3205, \text{ and } \sigma_\xi = 2.0786$$

$$\mu_{3\xi} = \frac{\Sigma\xi^3}{N} - 3\ M_\xi \frac{\Sigma\xi^2}{N} + 2\ M_\xi^3 = 2.7024$$

$$\mu_{4\xi} = \frac{\Sigma\xi^4}{N} - 4\ M_\xi \frac{\Sigma\xi^3}{N} + 6\ M_\xi^2 \frac{\Sigma\xi^2}{N} - 3\ M_\xi^4 = 87.7375$$

Since the interval = 3° we obtain

M = Arb. or. + iM_ξ = 81.7742, which, as explained later, should be published as 81.8.

$V = i^2\ V\xi = 38.8845$, published as 39.

$\sigma = i\ \sigma_\xi = 6.2358$, published as 6.2.

$\mu_3 = i^3 \mu_{3\xi} = 72.9648$, published as 7×10.

$\mu_4 = i^4 \mu_{4\xi} = 7106.7375$, published as 7.1×10^3.

It is seen that this method of moments, using grouped data and an arbitrary origin, has involved quite simple arithmetic. Also the work involved in getting the variance is directly contributive in getting the higher moments. Though this method is universally serviceable, some prefer a method based upon summations, for adding machines are more frequently available than multiplying machines. We illustrate a summation method, Elderton (1905-6), in Table VI F which, though involving larger figures than the method of moments, is applicable to grouped data, involves positive summations only, provides a check upon all summations as values appear twice, involves only such additions with totals and

subtotals as can be printed on an adding machine tape, so that the large numbers involved have constituted little mental tax.

TABLE VI F

THE COMPUTATION OF M, V, μ_3, AND μ_4, BY MEANS OF SUMMATIONS

X	f	ξ'	$f\xi'$	S_1	S_2	S_3	S_4
99	2	11	22	22	22	22	22
96	2	10	20	42	64	86	108
93	0	9	0	42	106	192	300
90	1	8	8	50	156	348	648
87	6	7	42	92	248	596	1244
84	13	6	78	170	418	1014	2258
81	23	5	115	285	703	1717	3975
78	5	4	20	305	1008	2725	6700
75	6	3	18	323	1331	4056	10756
72	1	2	-2	325	1656	5712	16468
69	1	1	1	$326=S_1$	$1982=S_2$	$7694=S_3$	$24162=S_4$
66	2	0					
	$62=N$		326	1982	7694	24162	

A brief examination of Table VI F suffices to reveal the mode of computation. The various summations from the arbitrary origin (here 66) are immediately obtainable from S_1, S_2, S_3, and S_4, from relationships herewith

$$\Sigma\xi' = S_1 \qquad [6:51]$$

$$\Sigma\xi'^2 = S_2 \qquad [6:52]$$

$$\Sigma\xi'^3 = 2\ S_3 - S_2 \qquad [6:53]$$

$$\Sigma\xi'^4 = 6\ S_4 - 6\ S_3 + S_2 \qquad [6:54]$$

Having the various sums of powers of ξ' substitution in [6:43], [6:44], [6:46], [6:47], and [6:48] yields the identical statistics already obtained by the method of moments, Table VI E, The summation method is well adapted to Hollerith and Powers machine computation.

A caution is necessary in using this method, or any method, involving large positive and negative quantities which nearly cancel. For example,

$$\mu_{3\xi'} = 216.2258 - 504.2653 + 290.7420 = 2.7023,$$

which is a five-figure answer though the addends are of seven figures. To secure a computational accuracy to five figures in the final answer has required computational accuracy to seven figures in the parts.

SECTION 7. DECIMAL PLACES TO BE KEPT IN PUBLISHED RESULTS

It is absurd to publish as many figures in final answers as have been here employed in the illustrative computations. The means of the first three temperatures of Table IV B, 80°, 88°, and 74°, is 80.666666. . . ad infinitum. The best evidence that we have from these three observations as to the population mean temperature is this answer kept to an infinite number of decimal places. Any rounding off whatsoever, e.g., 89.6666667, constitutes a deviation from the best possible answer yielded by the data. If, however, the confidence to be placed in the best answer (which is very small because only three observations have been utilized) differs from that to be placed in a rounded-off answer by a scarcely sensible amount, the rounded-off answer serves every purpose. Furthermore, if the rounding-off is done just at the decimal place, where confidence should flag, the rounded-off answer is more informative than the non-abridged answer. Following this principle, the mean of

the three measures should be published as 81°, implying that there is doubt as to the trustworthiness of the units figure, i.e., of the last figure published.

There is serious doubt as to any magnitude as small as one-third a standard error,—the chances of a magnitude as great as this arising merely as a matter of chance are 74 in 100. *We therefore adopt* as a practical rule *the practice of terminating a published statistic with the decimal place given by the first figure of one-third its standard error.* This rule, which we shall call the "one-third sigma rule", is equally applicable to original observations and to derived statistics. To apply it we must estimate standard errors where we cannot compute them. For all the more common and more important generalizing statistics formulas giving standard errors are available.

SECTION 8. **VARIANCE ERRORS AND STANDARD ERRORS OF MOMENTS**

Since a standard error is the square root of a variance error, it suffices to give formulas for variance errors.

Since the variance of the variance is $(\overline{V^2}-\bar{V}^2)$, we may utilize [6:26] and [6:10] and obtain

$$V_v = \frac{N-1}{N^3}(N-1)\tilde{U}_4 - (N-3)\tilde{V}^2 \qquad \text{[6:55]}$$

Variance of variance (samples of any size and parent populations of any form)

Since the sample μ_4 and V^2 are unbiased estimates of $\bar{\mu}_4$ and $\overline{V^2}$ and not of $\tilde{\mu}_4$ and $\tilde{V}^2$, an answer in terms of $\bar{\mu}_4$ and $\overline{V^2}$, for which we can substitute μ_4 and V^2 without introducing any

systematic bias, is desirable. We can derive such a formula by utilizing [6:10], [6:26], and [6:28]. The final result is

$$V_v = \frac{1}{N(N-2)(N-3)}\left[(N-1)^2\mu_4 - (N^2-3)\,V^2\right]$$

Variance error or sample variance, samples of any size and parent population of any form [6:56]

When N is large this reduces to

$$V_v = \frac{1}{N}(\mu_4 - V^2)$$

Variance error of sample variance, N large [6:57]

When the parent population is normal [6:56] reduces to

$$V_v = \frac{2V^2}{N+1}$$

Variance of variance when population is normal see (13:147) [6:58]

In Chapter XIII, Section 8, the variance of the standard deviation is derived. It is

$$V_\sigma = \frac{V_v}{4V}$$

Variance of σ, N large [6:59]

When N is large and the population normal this leads to

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$$

Standard error of σ, N large and population normal [6:60]

Small sample formulas for V_{μ_3} and V_{μ_4} become complicated, but the following given by Pearson (1902 – 1903 and 1914), are frequently serviceable

$$V_{\mu_3} = \frac{1}{N}(\tilde{\mu}_6 - \tilde{\mu}_3^2 + 9\,\tilde{V}^3 - 6\,\tilde{V}\,\tilde{\mu}_4)$$

Variance of μ_3, N large [6:61]

From which we obtain, approximately,

$$V_{\mu_3} = \frac{6}{N} V^3$$ Variance error of third moment,—N large and population normal [6:62]

$$V_{\mu_4} = \frac{1}{N}(\tilde{\mu}_8 - \tilde{\mu}_4^2 + 16\tilde{V}\,\tilde{\mu}_3^2 - 8\tilde{\mu}_3\tilde{\mu}_5)$$

Variance of μ_4,—N large [6:63]

From which we obtain, approximately,

$$V_{\mu_4} = \frac{96}{N} V^4$$ Variance error of fourth moment,—N large and population normal [6:64]

We may use these formulas to estimate the variance errors, and, by extracting the square roots, the standard errors, of the several statistics computed from the data of Table VI E, or Table VI F.

By [6:18] we find that $\sigma_m = .79$ and since one-third of this = .3, the mean is published as 81.8, indicating that the .8 is in doubt though it is better than a random guess.

By [6:56], which is the master or most trustworthy formula we have for V_v, we find $V_v = 94$ or $\sigma_v = 9.7$. Also, utilizing [6:60] we find $\sigma_\sigma = .78$. We accordingly publish V as 39. and σ as 6.2.

It is informative to see what the approximate formulas [6:58] and [6:60] yield. This latter is an important formula in that it does not involve the third or fourth moments and is widely used. By [6:58] we obtain, to two figures, the same answers as before, suggesting that [6:58] is satisfactory if N is as large as 62. Knowledge of the first significant figure of a standard error usually suffices to reveal the trust that can be placed in a statistic. Formula [6:60] yields $\sigma_\sigma = .56$. This value, though definitely

too small (because the assumption of normality was unsound), is nevertheless much to be preferred to no estimate at all as to the trustworthiness of σ.

Not having fifth, sixth, and eighth moments we cannot substitute sample moments for the unknown population moments and compute the variance errors of the third and fourth moments by the more accurate formulas [6:61] and [6:63]. We shall use the short formulas [6:62] and [6:64] knowing that resulting values will be much too small. By [6:62] we obtain $V_{\mu_3} = 5783.$ (to be published as $6. \times 10^3$, thereby revealing uncertainty as to the first figure), $\sigma_{\mu_3} = 76.$ $(= 8. \times 10)$, and $\sigma_{\mu_3}/3 = 25.$ Though this standard error is undoubtedly much too small, it is still probable that the first figure of μ_3 is better than a guess, so we publish it as $7. \times 10.$ By [6:64] we obtain $V_{\mu_4} = 3597885.$ $(= 4. \times 10^6)$, $\sigma_{\mu_4} = 1897$ $(= 2. \times 10^3)$, and $\sigma_{\mu_4}/3 = 632$ $(= 6 \times 10^2$. Though this value be too small, it may be that the first two figures of μ_4 have some merit, so we publish it as 7.1×10^3.

The unreliability here found for the higher moments is somewhat larger for these leptokurtic data than would be the case with a sample of equal size drawn from a normal population, but in general the higher moments are unreliable, so where possible crucial issues should be made to depend from statistics that do not involve them.

With reference to the reliability of estimates of the population variance and standard deviation, we note that since $\tilde{\tilde{V}} = [N/(N-1)]V$ and $\tilde{\tilde{\sigma}} = \sqrt{N/(N-1)}\ \sigma$ *the standard error of* $\tilde{\tilde{V}}$ *can be*

gotten by multiplying that of $\bar{V}$ *by* $N/(N-1)$ *and the standard error of* $\tilde{\sigma}$ *can be gotten by multiplying that of* σ *by* $\sqrt{N/(N-1)}$.

Treating the mean, variance, third and fourth moments in the same general situation has enabled a certain survey of types of statistics. Though we shall return again to the most important type, measures of simple variability of distributions and of derived statistics, the student should now have the basic concepts of variability and in particular of standard error, which will enable him to appraise the different measures of central tendency treated of in Chapter VII.

SECTION 9. THE AVERAGE DEVIATION

Though the standard deviation (or its square, the variance) is the most serviceable measure of dispersion because (1) of the importance and simplicity of the algebraic relationships in which it is found and because (2) it is generally the most reliable, there are many other measures, some of which have genuine merit for one reason or another.

The average deviation is nearly as reliable, in most situations, as the standard deviation, and as it is more readily computed it may be serviceable if the data are so extensive that the labor of computation is a decisive factor. It lacks the simplicity and properties that endear the standard deviation to the statistician, so when used it should be as a terminal statistic, —one not used in further combinatorial processes.

The average deviation is defined as *the arithmetic mean of the absolute values of the deviations of the measures in a series from their mean.* Thus

$$\text{A. D.} = \frac{\Sigma|X-M|}{N} \qquad \text{Definition of Average Deviation} \qquad [6:65]$$

If the absolute values of deviations are taken from some point, P, other than the mean, the resulting value must be labeled "average deviation from the point P."

$$\text{A.D. from } P = \frac{\Sigma|X-P|}{N} \quad \ldots\ldots\ldots\ldots\ldots \quad [6:66]$$

We have

$$\Sigma|X-P| = \overset{X\leq P}{\Sigma}(P-X) + \overset{X\geq P}{\Sigma}(X-P)$$

The superscript "$X<P$" indicates that the summation covers all values for which X is less than P. Values of $X \equiv P$ may be thought of as located in either of the right hand member summations, for these null differences will not affect the outcome.

$$\begin{aligned} \Sigma|X-P| &= \overset{X\leq P}{\Sigma}(P-X) + \overset{X\leq P}{\Sigma}(P-X) + \overset{X\leq P}{\Sigma}(X-P) + \overset{X\geq P}{\Sigma}(X-P) \\ &= 2\overset{X<P}{\Sigma}(P-X) + \Sigma(X-P) \\ &= 2\overset{X<P}{\Sigma}(P-X) + NM - NP \end{aligned}$$

If the number of measures below P is b, this may be written

$$\Sigma|X-P| = (2b-N)P + NM - 2\overset{X\leq P}{\Sigma}X$$

so that

$$\text{A. D. from } P = \frac{1}{N}[(2b-N)P + \Sigma X - 2\overset{X\leq P}{\Sigma}X] [6:67]$$

If measures are arranged in a frequency distribution and their sum, ΣX, gotten on an adding machine, it is only necessary to get the subtotal at the appropriate point to have $\overset{X\leq P}{\Sigma}X$. The average deviation from P thus is a very simple statistic to compute. If the A. D. is desired, a single listing, with a few sub-totals

in the neighborhood of the anticipated mean, will provide the two summations needed to yield both the mean and the average deviation. When $P = M$, formula [6:67] becomes

$$\text{A. D.} = \frac{2}{N}\left(bM - \sum^{X<M} X\right) \quad \text{The average deviation} \quad [6:68]$$

When P = the median formula [6:67] yields the average deviation from the median. This is a minimum. Otherwise expressed, *the median is such a point that the sum of the absolute values of the deviations from it of the measures in the series is a minimum.* The proof of this, which is very simple when all the measures in the series are different, is left as an exercise. To recapitulate: The basic and unique properties of the median are (1) the number of measures below it is equal to the number above it, and (2) the sum of the absolute deviations from it is a minimum.

The reliability of the average deviation. Since the average deviation from an arbitrary fixed point is simply the arithmetic mean of N absolute values, its standard deviation is given by the usual formula [6:18] for the standard error of the mean, using the absolute values (Y below) as the measures of the series.

Let X = the original measures of the series.
Let P = a fixed point, or value of X.
Let $Y = |X-P|$
Then A. D. from P = the mean $Y = \overline{Y}$. $y = Y - \overline{Y}$
and

$$\sigma_{\text{A. D. from } P} = \frac{\sigma_y}{\sqrt{N-1}} \quad \text{Standard deviation of average deviation from fixed point } P \quad [6:69]$$

If a point which is a linear function of the data, as is the mean, is the point of reference, one degree of freedom is consumed thereby. Thus we have for the standard deviation of the average

deviation:

Let X = the original measures of the series

Let $Z = |X-M|$

Then

A. D. = the mean $Z = \bar{Z}$

$$z = Z - \bar{Z}$$

and

$$\sigma_{A.D.} = \frac{\sigma_z}{\sqrt{N-2}} \quad \text{Standard deviation of the average deviation} \qquad [6:70]$$

Though in general the standard deviation is a more reliable measure than the average deviation, i.e., σ/σ_σ is greater than A.D./$\sigma_{A.D.}$, leptokurtic distributions exist for which this is not the case. In R. A. Fisher (1921) is a basic treatment of, "consistent" and "efficient" statistics.

SECTION 10. BASED UPON PERCENTILES

The distance between any two percentiles, $P_p - P_{p'}$, is an inter-percentile range and is a measure of variability, though a very poor measure if the two proportions determining the percent tiles are near together. It is also a poor measure if either of the proportions is very near zero or one, for the corresponding percentile is much less reliable in the case of unimodal distributions not arbitrarily bounded at an extreme than a percentile somewhat removed from zero or one hundred.

The most reliable interpercentile range is a function of the form of the population distribution. The standard error of $P_p - P_{p'}$ is derivable from the standard deviation of a percentile [4:03] and the covariance (as explained in Chapter X, Section 5) between percentiles. Using

c to indicate covariance, $p>p'$, and other symbols as earlier defined, it can be shown that

$$c\,(P_p P_{p'}) = N \frac{i\,q\;\; i'\,p'}{f_p \;\; f_{p'}} \quad \text{Covariance between } P_p \text{ and } P_{p'} \text{ percentiles} \qquad [6:71]$$

Utilizing [4:03] and the formula for the variance of a difference [10:106] we obtain

$$V(P_p - P_{p'}) = N\left(\frac{i^2 p\, q}{f_p^2} + \frac{i'^2 p'\, q'}{f_{p'}^2} - \frac{2iqi'p'}{f_p \;\; f_{p'}}\right)$$

Variance of an interpercentile range [6:72]

The most reliable interpercentile range is that one for which $(P_p - P_{p'})/\sigma_{P_p - P_{p'}}$ is a maximum. This can be determined when the form of distribution is known, and the writer has shown (Kelley 1921), that, to a close approximation, this is, in the case of a normal distribution, *the distance from the 7th to the 93rd percentiles.* This *can be designated the normal optimal interpercentile range*, but since it has merit for many non-normal unimodal distributions such as are commonly found, *we will* give it a general designation and *call it Pv.*

$$Pv = P_{.93} - P_{.07} \quad \text{A meritorious percentile measure of variability} \qquad [6:73]$$

Letting $p = .93$ and $p' = q = .07$, the variance of Pv is given by [6:72] which then becomes

$$V_{Pv} = N\left(\frac{.0651\; i_p^2}{f_p^2} + \frac{.0651\; i_q^2}{f_q^2} - \frac{.0098\;\; i_p i_q}{f_p \;\; f_q}\right)$$

Variance of error of Pv [6:74]

A percentile measure of variability that has had wide usage is *the quartile deviation, or the semi-interquartile range,* which *is commonly designated Q.* The reader must distinguish be-

tween Q and Q_1, Q_2, and Q_3, which are common designations of the 1st, 2nd, and 3rd quartiles or $P_{.25}$, $P_{.50}$, and $P_{.75}$ respectively.

$$Q = \frac{P_{.75} - P_{.25}}{2}$$ Quartile deviation [6:75]

The variance error of Q is given by [6:76]

$$V_q = \frac{N}{4}\left(\frac{.1875\, i^2}{f_p^2} + \frac{.1875\, i'^2}{f_{p'}^2} - \frac{.125\, i\, i'}{f_p\, f_{p'}}\right)$$

Variance error of quartile deviation [6:76]

SECTION 11. **COMPARISON OF SUNDRY MEASURES OF VARIABILITY**

The relative standard error of Q is generally much larger than that of Pv, which in turn is generally larger than that of A.D., which in turn is generally larger than that of σ. The relationships of these measures of variability in a normal distribution are shown in Table VI G.

TABLE VI G

RATIOS OF CERTAIN MEASURES OF VARIABILITY TO THEIR STANDARD ERRORS IN THE CASE OF SAMPLES DRAWN FROM A NORMAL POPULATION

STANDARD ERROR FORMULAS	CRITICAL RATIOS	SIZE OF SAMPLE TO YIELD RELATIVE RELIABILITY EQUAL TO THAT OF σ FROM A SAMPLE OF 100
[6:60]*	$\frac{\sigma}{\sigma_\sigma} = 1.4142\sqrt{N}$	100
[6:70]*	$\frac{\text{A.D.}}{\sigma_{\text{A.D.}}} = 1.3236\sqrt{N}$	114
[6:74]*	$\frac{Pv}{\sigma_{Pv}} = 1.1421\sqrt{N}$	153
[6:76]*	$\frac{Q}{\sigma_q} = .8573\sqrt{N}$	272

* Formulas used, except as modified to fit theoretical normal distributions.

Table VI G shows that the average deviation is nearly as reliable as the standard deviation, that the quartile deviation is a very unreliable measure, and that *Pv*, though the best of the interpercentile measures, requires 53 per cent more cases than the standard deviation to obtain comparable reliability in the case of samples drawn from a normal population. The formula for the variance of an interpercentile range [6:72] reveals that the 10th to 90th percentile range is nearly as reliable as *Pv* and may well be used in lieu thereof in situations wherein the 10th and 90th percentiles are already available.

Where considerations of simplicity or immediate meaningfulness of statistic has led to the use of percentiles it is consistent to employ an interpercentile measure as a measure of variability, and either *Pv* or the 10 to 90 percentile range is recommended. In this same situation percentile measures of skewness and of kurtosis, as given in Chapter VII are recommended, but the statistician will realize that all these percentile measures are terminal statistics and very cumbersome to incorporate into further computational procedures.

CHAPTER VII

MEASURES OF CENTRAL TENDENCY

SECTION 1. THE EFFECT OF FORM OF DISTRIBUTION UPON DIFFERENT AVERAGES

Averages: In common usage the term "average" indicates the sum of the measures divided by their number, but in statistical parlance this is the "arithmetic mean," and any measure whatsoever of central tendency is an "average." According to a common definition any function of a series of measures is an average of them if it equals them in the special case when they all have the same value. This definition is so broad as not in itself to guarantee useful properties, so we will consider a class of averages which are more restricted.

If X_1, X_2, X_3 . . . X_N are all positive and are the measures in the series, a general expression giving many averages, but not all possible ones, is

$$f(b) = \left\{ \frac{\Sigma X^b}{N} \right\}^{1/b} \quad . . . \quad \text{A Generalized Mean} \qquad [7:01]$$

As b takes different values we obtain different sorts of means. When b=1, equation [7:01] yields

the most common mean, the arithmetic mean. When $b=-1$, equation [7:01] yields the harmonic mean. When $b=2$, equation [7:01] yields the quadratic mean. This is not the standard deviation because ΣX does not $= 0$. When $b=0$, equation [7:01] yields an indeterminate form, which can be shown to be the geometric mean.*

* By definition we have the following for the geometric mean

$$\text{G.M.} = (X_1 X_2 X_3 \ldots X_N)^{1/N}$$

or

$$\log \text{G.M.} = \frac{\Sigma(\log X)}{N}$$

we also have

$$[f(b)]^b = \frac{\Sigma X^b}{N} = \frac{X_1^b + X_2^b + X_3^b \ldots + X_N^b}{N}$$

Since X^b can be expanded into a convergent series, as follows:

$$X^b = 1 + b \log X + b^2 \frac{(\log X)^2}{2!} + \text{higher powers in } b$$

We can set

$$X^b = 1 + b \log X, \quad \text{as } b \text{ approaches zero.}$$

Thus

$$[f(b)]^b = \frac{1}{N}[1+b \log X_1 + 1 + b \log X_2 + \cdots$$
$$+ 1 + b \log X_N]$$
$$= 1 + b \frac{\Sigma(\log X)}{N} = 1 + b \log \text{G.M.} = (\text{G.M.})^b$$

Giving

$$f(b) = \text{G.M.}, \quad \text{as } b \text{ approaches zero.}$$

When $b = -\infty$, equation [7:01] yields the smallest measure in the series and when $b = \infty$, it yields the largest measure. In all cases $f(b)$ increases as b increases from $-\infty$ to ∞, immediately proving that all of the means given by [7:01] are internal means, i.e., they fall within the range of the measures in the series, and also proving that the harmonic mean is smaller than the geometric mean, which is smaller than the arithmetic mean, which is smaller than the quadratic mean, etc. If reasons inherent in, or exterior to, the data assert that small measures should play a more important part than large measures in determining the average, it follows that b of equation [7:01] should, say, equal 0 or -1, rather than 1,—i.e., that the geometric or the harmonic mean should be used rather than the arithmetic mean. Averages other than those defined by [7:01] exist, such as the median and the mode.

The averages that we shall consider in some detail are the median, the mode, the geometric mean, the harmonic mean, and the arithmetic mean, which last will frequently be called "the mean."

The usual order of trustworthiness of various measures of a distribution: For many distributions of quantitative data the features of greatest stability are, in order:

1. The number of cases, N. This is usually fixed by the experimenter and thus has no sampling or experimental error. If the principle of sampling does not fix the number of cases, as would be the case if the ages of all the men passing a selected spot in a certain hour were recorded, then of course there is a sampling error in N, but this is not the usual mode of sampling. We may therefore generally think of the number of cases in the sample as having no sampling error.

2. Some measure of variability, generally the

standard deviation. Contrary to the usual untutored expectation which would place the mean first, it is generally a fact that a measure of variability has greater reliability.

3. Some average, generally the arithmetic mean.

4. Some measure of skewness, such as μ_3/σ_{μ_3} (see [6.47], [6:61], and [6:62]), As/σ_{As} (see [7:16] and [7:17]), β_1/σ_{β_1} (see K. Pearson, *Tables*), Sk/σ_{sk} (see [7:18] and [7:19]), or g_1/σ_{g_1} (see R. A. Fisher, 1934).

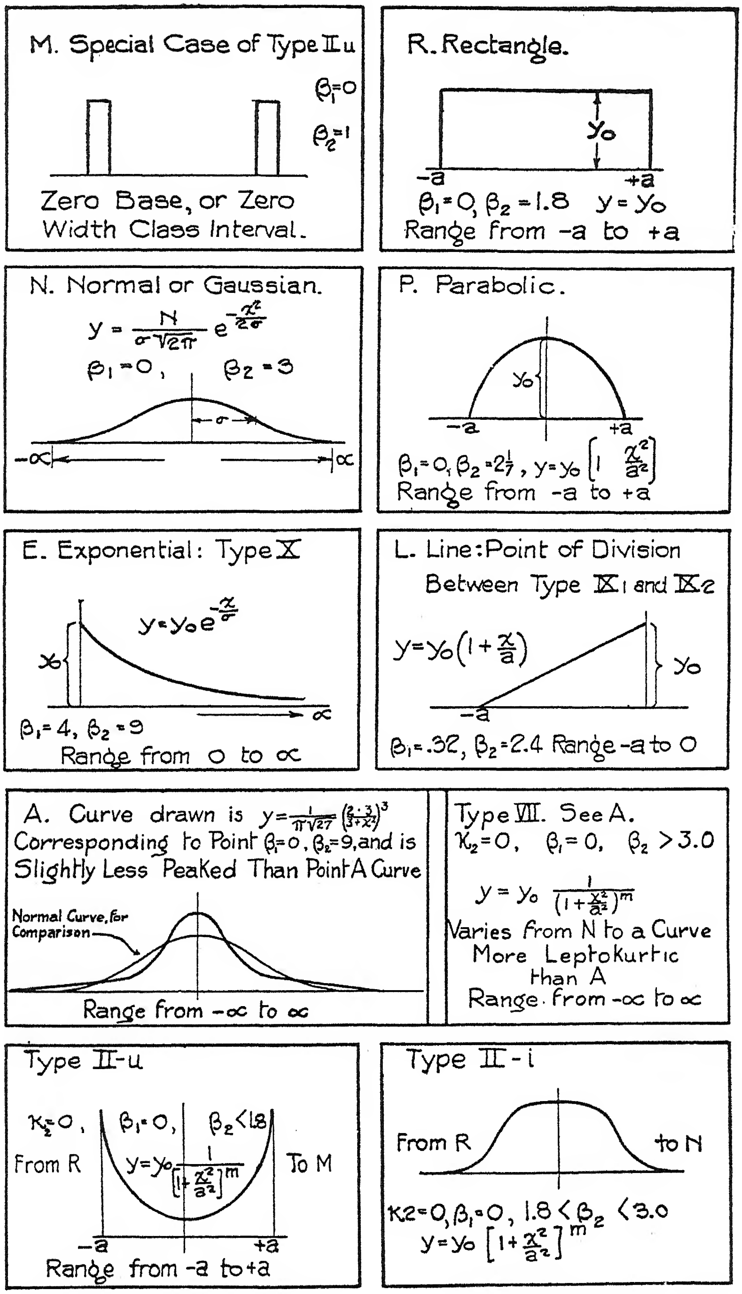
5. Some measure of kurtosis, such as $(\beta_2-3)/\sigma_{\beta_2}$ (see K. Pearson, *Tables*), Ku-$2.7885/\sigma_{Ku}$ (see, [7:08] and [7:09]), g_2/σ_{g_2} (see R. A. Fisher, 1934).

6. Some measures of multimodality.

The order of reliability of statistics here given is not invariable for it is affected by the form of distribution of the data. Certain unusual forms may augment the reliability of some of the later measures in this list, or of other measures not mentioned, while decreasing that of some of the earlier measures. Also the list is suspect because the basis of comparison of the disparate measures has not been given and defended. Is the basis that of comparing percentage errors, or absolute errors, or something else? A defense based upon statistical concepts thus far presented in this text is impossible, so it must suffice to ask the reader to return to this issue of the order of trustworthiness of different measures as his knowledge about them increases. Until he can form an independent judgment, based upon his particular data, he is advised to attempt so to set up his issues that

statistics low rather than high in this list provide the requisite information. For example, suppose the question is, "Are there more men than women of a high level of intelligence?" One study might compare averages of men and women, another compare variabilities of men and of women, and another investigate the kurtosis of men and of women. Since each of these bear upon the issue and since variability is number two in the list, the average number three, and kurtosis number five, the initial approach suggested is via a study of variabilities of men and women. Certainly a study based upon proof of multimodality should be undertaken as a sort of last resort, that is, only in case no measures lower in the list will answer the issue.

Common forms of distribution: Charts VII I, VII II, VII III, show various Pearson-type curves. The equations of these curves will not now concern us. Where two curves are shown under a single-type number, one of them is a rather typical curve of this type and the other an extreme curve in one direction of the type in question, and the letter near the left or right margin gives the extreme curve in the other direction. For example, the two category type varies from the *M* (for Mendelian, because of the importance in heredity studies of the two-point distribution having equal frequencies in the two categories) curve wherein are equal frequencies in two classes through the curve having unequal frequencies in two classes to the limiting case where all the frequencies lie in one class. The letters *M*, *R* (rectangular), *N* (normal), *P* (parabolic), *E* (exponential), and *L* (straight line) stand for curves as illustrated in the first six figures of Chart VII I. All curves having a single mode not at a dictated upper or lower boundary are referred to as unimodal, or "*i*", or "Λ", curves; all having an anti-mode are called


M. Special Case of Type II u
$\beta_1=0$
$\beta_2=1$
Zero Base, or Zero Width Class Interval.
R. Rectangle.
y_0
$-a$
$+a$
$\beta_1=0,\ \beta_2=1.8\quad y=y_0$
Range from $-a$ to $+a$
N. Normal or Gaussian.
$y=\frac{N}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma}}$
$\beta_1=0,\quad \beta_2=3$
σ
$-\infty$
∞
P. Parabolic.
y_0
$-a$
$+a$
$\beta_1=0,\ \beta_2=2\frac{1}{7},\ y=y_0\left[1\quad\frac{x^2}{a^2}\right]$
Range from $-a$ to $+a$
E. Exponential: Type X
$y=y_0e^{-\frac{x}{\sigma}}$
y_0
∞
$\beta_1=4,\ \beta_2=9$
Range from 0 to ∞
L. Line: Point of Division
Between Type IX 1 and IX 2
$y=y_0\left(1+\frac{x}{a}\right)$
y_0
$-a$
$\beta_1=.32,\ \beta_2=2.4$ Range $-a$ to 0
A. Curve drawn is $y=\frac{1}{\pi\sqrt{27}}\left(\frac{3\cdot 3}{3+x^2}\right)^3$
Corresponding to Point $\beta_1=0,\ \beta_2=9$, and is
Slightly Less Peaked Than Point A Curve
Normal Curve, for Comparison
Range from $-\infty$ to ∞
Type VII. See A.
$\kappa_2=0,\quad \beta_1=0,\quad \beta_2>3.0$
$y=y_0\frac{1}{\left(1+\frac{x^2}{a^2}\right)^m}$
Varies from N to a Curve
More Leptokurtic
than A
Range from $-\infty$ to ∞
Type II-u
$\kappa_2=0,\quad \beta_1=0,\quad \beta_2<1.8$
From R
$y=y_0\frac{1}{\left[1+\frac{x^2}{a^2}\right]^m}$
To M
$-a$
$+a$
Range from $-a$ to $+a$
Type II-i
From R
to N
$\kappa_2=0,\ \beta_1=0,\ 1.8<\beta_2<3.0$
$y=y_0\left[1+\frac{x^2}{a^2}\right]^m$

anti-modal, or "*u*", curves, and all having neither a mode nor anti-mode except at a boundary are called *J*-shaped, or "*J*", curves. These Pearson-type curves seem to cover pretty well the range of the simpler phenomena of distribution. One omission which should be classed as simple is the point distribution. This consists of frequencies at discrete points along a graduated scale. As an example may be mentioned the number of children in families,—the frequencies occur at 0, 1, 2, 3, etc., but never at fractional values. Many point distributions, especially those having ten or more classes, may be excellently handled as one would handle grouped data of a continuous variable. Certain cautions will be obvious. For example, we may speak of the mean number of children per family as 2.30, but we hardly should say that the median or the modal family has a fractional number of children. (See Elderton, 1927.)

SECTION 2. THE MEDIAN

The computation of the median, or the fiftieth percentile, $P_{.5}$, and of its standard error has been given in Chapter IV, Section 3. Formula [4:03] giving the standard error of any percentile is

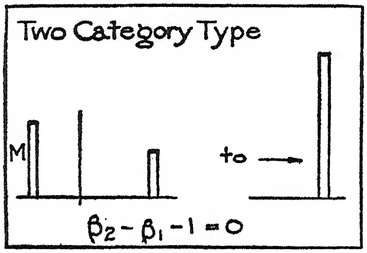
$$\sigma_{P_p} = \frac{i'_p \sqrt{Npq}}{f'_p} \quad \ldots\ldots\ \text{See [4:03]}$$

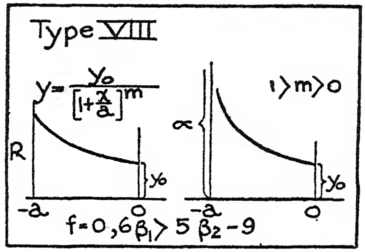
For the median $p = q = .5$ and this formula becomes

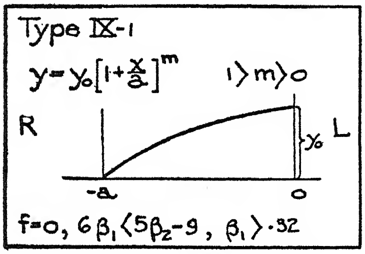
$$\sigma_{Mdn} = \frac{i'_p \sqrt{N}}{2f'_p} \quad \ldots\ldots \quad [7:02]$$

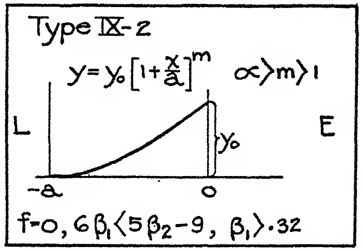
Equation [7:02] may be written

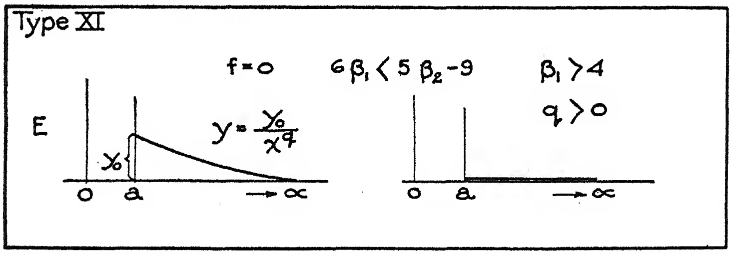
* Elderton's "twisted j-shaped" type is one in which the small tail of the j turns down instead of up.

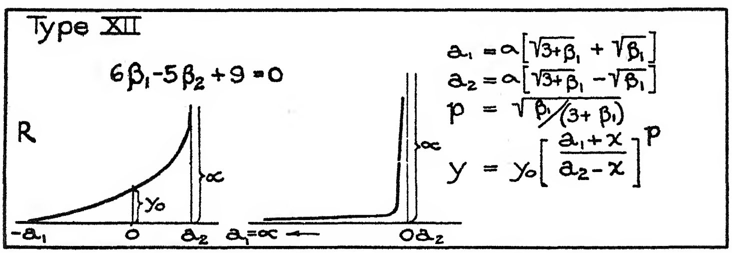

Two Category Type
M
to
$\beta_2-\beta_1-1=0$

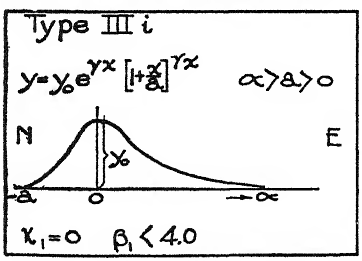

Type VIII
$y=\dfrac{y_0}{[1+\frac{x}{a}]^m}$
$1>m>0$
R
∞
y_0
$-a$ 0 $-a$ 0
$f=0, 6\beta_1>5\beta_2-9$

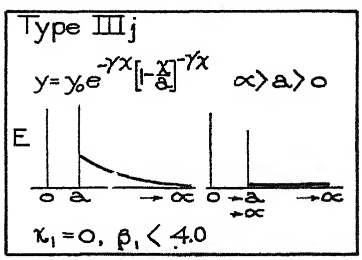

Type IX-1
$y=y_0[1+\frac{x}{a}]^m$
$1>m>0$
R
L
y_0
$-a$ 0
$f=0, 6\beta_1<5\beta_2-9, \beta_1>.32$


Type IX-2
$y=y_0[1+\frac{x}{a}]^m$
$\infty>m>1$
L
E
y_0
$-a$ 0
$f=0, 6\beta_1<5\beta_2-9, \beta_1>.32$


Type XI
$f=0$
$6\beta_1<5\beta_2-9$
$\beta_1>4$
$q>0$
E
$y=\dfrac{y_0}{x^q}$
y_0
0 a $\to\infty$
0 a $\to\infty$


Type XII
$6\beta_1-5\beta_2+9=0$
$a_1=\alpha[\sqrt{3+\beta_1}+\sqrt{\beta_1}]$
$a_2=\alpha[\sqrt{3+\beta_1}-\sqrt{\beta_1}]$
$p=\sqrt{\beta_1/(3+\beta_1)}$
$y=y_0\left[\dfrac{a_1+x}{a_2-x}\right]^p$
R
∞
y_0
∞
$-a_1$ 0 a_2 $a_1=\infty\leftarrow$
$0\,a_2$


Type III i
$y=y_0e^{\gamma x}[1+\frac{x}{a}]^{\gamma x}$
$\infty>a>0$
N
E
y_0
$-a$ 0 $\to\infty$
$\kappa_1=0$ $\beta_1<4.0$

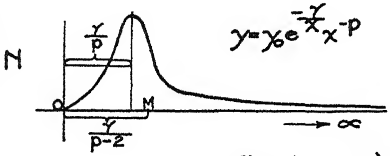

Type III j
$y=y_0e^{-\gamma x}[1-\frac{x}{a}]^{-\gamma x}$
$\infty>a>0$
E
0 a $\to\infty$ 0 $\to a$ $\to\infty$
$\to\infty$
$\kappa_1=0, \beta_1<4.0$

$$\sigma_{Mdn} = \frac{1}{\frac{2f'_p}{i'_pN}\sqrt{N}} \quad . \; . \; . \; . \qquad [7:02]$$

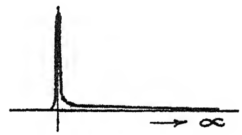
The quotient $\frac{f'_p}{i'_pN}$ is the proportionate density of frequency per unit interval, and as this is found in the denominator we see that the standard error of the median decreases as the density of cases increases. Accordingly, the median becomes a better and better measure as the density of cases in the neighborhood of the median increases. Though, generally speaking, the mean is more reliable than the median, this is not so if the distribution is sufficiently leptokurtic. Curve A, Chart VII I, is a symmetrical unimodal leptokurtic curve for which, approximately, the standard error of the median is equal to that of the mean. This is a curve having greater relative frequency both near the middle (which increases the reliability of the median) and near the extremes (which decreases the reliability of the mean) than does the normal curve. Another wording which applies to unimodal curves, is that *a leptokurtic curve is low at the hips*, positions intermediate between the central and the extreme portions. *A mesokurtic curve*, of which the normal curve is an example, *has medium hip height, and a platykurtic curve is high at the hips*.

The more leptokurtic the distribution the greater the merit of the median as an average, but we must note that a high degree of leptokurtosis is necessary before the standard error of the median becomes less than that of the mean. We will shortly provide measures of kurtosis, but even without them the reader can get a good visual impression of Curve A and then expect that if his data are still more leptokurtic the median is to be preferred to the mean as a measure of

Type V



$$y = y_0 e^{-\frac{\gamma}{x}} x^{-p}$$



$\kappa_2 = 1 \qquad p > 5$

Type I-u

Boundaries: Two Category. $\kappa_2 < 0$ Above $f = 0$

Type II-u
Type VIII

$$y = y_0\left(1+\frac{x}{a_1}\right)^{-m_1}\left(1-\frac{x}{a_2}\right)^{-m_2}$$

Type I-j

Boundaries: $f = 0$ $\kappa_2 < 0$ Inside $f = 0$

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$$

Type I-i

Boundaries: Types IX-1 $\kappa_2 < 0$ Below $f = 0$
IX-2
II-i
III-i

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$$

Type VI-i

$\infty > \kappa_2 > 1.00$ Below $f = 0$

Boundaries: Types III-i
XI
V

$$y = \frac{y_0 (x-a)^{m_1}}{x^{m_2}}$$

$m_1 > 0 \qquad m_2 > 0$

Type VI-j

$\infty > \kappa_2 > 1.00$ Inside $f = 0$

Boundaries: Types III-j
XI

$$y = y_0 \frac{(x-a)^{m_1}}{x^{m_2}}$$

$m_1 < 0 \qquad m_2 > 0$

Type IV

$1 > \kappa_2 > 0$

Boundaries: Types V
VII

$$y = y_0 \frac{e^{-\nu \tan^{-1}\frac{x}{a}}}{\left(1+\frac{x^2}{a^2}\right)^m}$$

central tendency.

For all distributions the simplicity of meaning of the median and of interpercentile ranges highly recommends them for use with popular audiences, but only in the case of highly leptokurtic distributions do they have superior status because of reliability.

An exact and general quantitative relationship between the kurtosis of a curve and the relative excellence of mean and median is impossible because distributions with the same kurtosis, [7:03], may be non-identical in other respects, such as the relative frequency in the neighborhood of the median.

$$\beta_2 = \frac{\mu_4}{V^2} \qquad \text{Pearson's measure of kurtosis} \qquad [7:03]$$

β_2 = 3.00 for mesokurtic distributions, of which the normal is an example. Thus for a test of divergence from mesokurtosis we have

$$\frac{\beta_2 - 3}{\sigma_{\beta_2}} \qquad \text{Critical ratio, providing mesokurtosis test based upon } \beta_2 \qquad [7:04]$$

Pearson (1914) has provided a table from which an approximate and entirely serviceable value of σ_{β_2} can be found.

The obvious and crucial test of excellence of mean and median is to compute the standard errors of each for the distribution in question and assert that the one with the smaller standard error is the more reliable. Some practice in doing this will provide the student with such insight that he can generally tell which is the more reliable measure by looking at the graph of the distribution. Equations [7:05] and [7:06] following are of two leptokurtic distributions

having equally reliable mean and median. In appearance each deviates slightly from Curve A of Chart VI I. Curve [7:05] is one given by the sum of two normal distributions and [7:06] is a Pearson Type VII curve.

If in an equation e has an exponent, $f(x)$, which is elaborate, it makes for typographical simplicity to write "exp $f(x)$" in lieu of "$e^{f(x)}$", as has been done in [7:05]:

$$y = \frac{N}{2\sigma\sqrt{2\pi}}\left(k_1 \exp \frac{-k_1^2 x^2}{2\sigma^2} + k_2 \exp \frac{-k_2^2 x^2}{2\sigma^2}\right) \quad [7:05]$$

in which k_1 and k_2 the roots of

$$k = \sqrt{.5\pi} \pm \sqrt{.5\pi + .5 - \sqrt{.25 + \pi}} = 1.7320 \text{ or } .7746.$$

The fourth moment of [7:05] is

$$\mu_4 = \frac{24\,[4\pi^2 + 6\pi + 1 - (4\pi - 1)\sqrt{1 + 4\pi}]\ \sigma^4}{(1 - \sqrt{1 + 4\pi})^4} = 4.3333\sigma^4$$

Thus for this curve $\beta_2 = 4.3333$.

$$y = \frac{N}{2\left(1 + \frac{x^2}{2.6862\,\sigma^2}\right)^{2.8431}} \quad \ldots\ldots\ldots \quad [7:06]$$

The Pearson β_2 is 11.7434, but the standard error of β_2, as computed by Pearson's method, for this very leptokurtic Type VII curve is infinite. We may use the following as a preliminary guide, but not as a substitute for the crucial test mentioned:

$\beta_2 > 5$ β_2 standard suggesting the median as a more reliable measure than the mean [7:07]

A similar standard, based upon percentiles, and one having a finite standard error even with extreme leptokurtosis, may be based upon the quotient of certain interpercentile ranges. Trial suggests that for this purpose the quotient between a greater and a lesser interpercentile range than *Pv* will serve. Calling the 3 to 97 interpercentile range *Tns* (initials of "three-ninety-seven") we have

$$Ku = \frac{P_{.97} - P_{.03}}{P_{.25} - P_{.25}} = \frac{Tns}{2Q} \qquad \text{Percentile measure of kurtosis} \qquad [7:08]$$

The variance error of *Ku* is obtainable from [7:09] in which $p = .97$ and $p' = .75$.

$$V(Ku) = N\ Ku^2 \Big[\frac{i_p i_q pq}{f_p^2\ Tns^2} + \frac{i_p\ i_q\ pq}{f_q^2\ Tns^2} - \frac{2i_p\ i_q\ q^2}{f_p\ f_q\ Tns^2}$$

$$+ \frac{i_{p'} i_{q'} p'q'}{4f_{p'}^2 Q^2} + \frac{i_{p'}\ i_{q'}\ p'q'}{4f_{q'}^2 Q^2} - \frac{i_{p'}\ i_{q'}\ q'^2}{4f_{p'}\ f_{q'}\ Q^2}$$

$$- \frac{i_p\ i_{q'}\ qq'}{2QTns f_p\ f_{q'}} - \frac{i_q\ i_{p'}\ qq'}{2QTns f_q\ f_{p'}} + \frac{i_p\ i_{p'}\ qp'}{2QTns f_p\ f_{p'}}$$

$$+ \frac{i_q\ i_{q'}\ qp'}{2QTns f_q\ f_{q'}}\Big] \qquad \text{Variance error of Ku} \qquad [7:09]$$

For [7:05] $Ku = 4.60$ and its standard error $= 8.27/\sqrt{N}$. We may use the following as a guide:

$$Ku > 5 \qquad \text{Ku standard suggesting the median as a more reliable measure than the mean} \qquad [7:10]$$

In a normal distribution

$$Ku = 2.7885 \qquad \text{Mesokurtic Ku} \ldots\ldots\ldots\ldots \qquad [7:11]$$

In view of the extreme unreliability of higher moments, and even of the standard deviation, *in highly leptokurtic distributions* we may conclude that when the *median is preferred to the mean an interpercentile measure of variability, such as Pv* [6:73], *is to be preferred to the standard deviation, a percentile measure of skewness, such as Sk* [7:18], *is to be preferred to* β_1 [7:12], *and a percentile measure of kurtosis, such as Ku* [7:08], *is to be preferred to* β_2 [7:03].

The utility of measures of skewness is discussed in Chapter VII, 3, in connection with the mode.

In certain problems the selection of measures because of high reliability may have to yield to selection because of amenability to algebraic manipulation. This latter consideration may dictate the use of M, V, μ_3, μ_4, β_1, β_2, or of

Fisher's unbiased k-statistics, $k_1 (=M)$, k_2 $(=\tilde{\bar{V}})$,

$$k_3(=\bar{\tilde{\mu}}_3),\ k_4\left\{=\frac{N^2}{(N-1)(N-2)(N-3)}+[(N+1)\mu_4-3(N-1)V^2]\right\},$$

$$g_1\ (g_1^2=\frac{k_3^2}{k_2^3}),\ \text{and}\ g_2\ (=\frac{k_4}{k_2^2}).$$

Summarizing the points noted about the median, we have:

(1) Its simple and direct meaning recommends its use with popular audiences.

(2) Its standard error can be simply calculated.

(3) It is a terminal statistic and seldom should be used in further computation for its algebraic manipulation is difficult.

(4) Its reliability increases as the kurtosis of the distribution increases, becoming greater than that of the mean for pronouncedly leptokurtic distributions.

(5) Though it does not lose its simple meaning in the case of skewed distributions, its greatest utility as a measure of central tendency is with distributions which are approximately symmetrical.

SECTION 3. THE MODE

For distributions which are nearly symmetrical the practical issue generally is whether the measure of central tendency shall be the mean or the median, or whether both shall be used, and for skewed distributions it is whether it shall be the mode, the harmonic mean, or the geometric mean, or whether one of these and the mean shall be used. The mean holds a preferred position in both of these general situations because of its nice algebraic properties.

The harmonic and geometric means are limited to distributions, generally skewed, wherein zero and negative values of the variate are impossible. No such limitation attaches to the mean, median, or mode, but the peculiar merit of the mode as the measure of maximum likelihood is most apparent and informative in the case of skewed distributions only. We therefore discuss the mode in connection with skewness.

Measures of asymmetry and of skewness: *We will* in this text *call a measure of imbalance of a distribution a measure of asymmetry if it involves the units of measurement and a measure of skewness if it is independent of them.* Thus μ_3 is a measure of asymmetry and μ_3/σ^3 is a measure of skewness, for the former does and the latter does not change as the units of measurement are multiplied or divided by any constant. Pearson's

measure of skewness:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \text{Pearson's } \beta_1 \text{ measure of skewness *} \qquad [7:12]$$

Tables yielding the standard error of β_1 for different-sized samples and different values of β_1 and β_2 are given in Pearson's *Tables for Statisticians and Biometricians*.

Fisher has devised a very similiar measure of skewness based, not upon the sample variance and third moment, but upon unbiased estimates of the population variance and third moment. His measure is

$$g_1 = \frac{k_3}{\sqrt{k_2^3}} \qquad \text{Fisher's measure of skewness} \qquad [7:13]$$

A general formula for the variance error of g_1 is not available, though frequently needed when dealing with clearly skewed distributions. Fisher (Statistical Methods for Research Workers, 1925 et seq.) does provide the variance error in the case of a sample drawn from a normal population and this, of course, is all that is needed to test departure from normality. It is

$$V_{g_1} = \frac{6N\ (N-1)}{(N-2)\ (N+1)\ (N+3)} \qquad \text{Variance error of } g_1 \text{ in case of sample drawn from a normal population} \qquad [7:14]$$

This formula is recommended not only by its simplicity, but because it applies to small as well as large samples.

The critical ratios μ_3/σ_{μ_3} and k_3/σ_{k_3} are serviceable as tests of asymmetry. Formulas

* Not to be confused with Pearson's $(Mo-M)/\sigma$ which he designates as "skewness",—the mode herein having been determined from a fitted curve.

[6:61] and [6:62] give the variance error of μ_3 in case N is large. In the special case in which N is large and the population normal it can readily be shown that

$$V_{k_3} = \frac{6k_2^3}{N} = \frac{6V^3}{N} \quad \text{Variance error of } k_3\text{, N large and population normal} \qquad [7:15]$$

The Pearson β_1 and Fisher g_1 measures of skewness suffer from extreme unreliability for ordinary-sized samples whenever, as is quite common in economic and psychological data, skewness is associated with considerable leptokurtosis. In these situations the simpler tests of asymmetry, μ_3/σ_{μ_3} and k_3/σ_{k_3}, suffer for the same reason. To serve with data of this skewed leptokurtic sort we give the following percentile measures of asymmetry and skewness:

$$As = \frac{P_{.07} + P_{.93}}{2} - P_{.50} \quad \text{Percentile measure of asymmetry} \qquad [7:16]$$

The variance error of this measure is given herewith, wherein p = .93 and p' = .50.

$$V_{As} = \frac{N}{4}\left[pq\left(\frac{i_p^2}{f_p^2} + \frac{i_q^2}{f_q^2}\right) + \frac{2i_p i_q q^2}{f_p\ f_q} + \frac{i_{p'}^2}{f_{p'}^2} - \frac{2qi_{p'}}{f_{p'}}\left(\frac{i_p}{f_p} + \frac{i_q}{f_q}\right)\right]$$

Variance of percentile measure of asymmetry [7:17]

For an abstract measure of skewness we have

$$Sk = \frac{As}{Pv} \quad \text{Percentile measure of skewness} \qquad [7:18]$$

The variance error of this measure is given herewith

$$V_{Sk} = \frac{1}{Pv^4}\left\{Pv^2 V(As) + As^2 V(Pv) + Pv\ As\ \left[pq\left(\frac{i_p^2}{f_p^2} - \frac{i_q^2}{f_q^2}\right) + 2qq'\frac{i_{p'}}{f_{p'}}\left(\frac{i_q}{f_q} - \frac{i_p}{f_p}\right)\right]\right\} \quad \text{Variance of percentile measure of skewness} \qquad [7:19]$$

Except in the case of β_1, which is always positive, the sign of the other measures, μ_3, μ_3/σ^3, k_3, g_1, *As*, and *Sk*, is positive and the skewness is called positive when the long tail of the distribution is to the right. i.e., toward high values of the variate. The expression "skewed to the right" has been varicusly used. It can easily be avoided, but if used it is recommended that it be synonymous with "positive skewness" as here defined.

We will illustrate these several measures in connection with the distribution of Table VII A.

TABLE VII A

RELATIVE FREQUENCIES IN A PEARSON TYPE III, $\beta_1 = 1$, $\beta_2 = 4.5$, DISTRIBUTION, FOR $.2\ \sigma$ INTERVALS

x	f	x	f	x	f
Up to					
-1.8	.0008	.5	.0562	2.9	.0035
-1.7	.0083	.7	.0474	3.1	.0027
-1.5	.0247	.9	.0394	3.3	.0020
-1.3	.0450	1.1	.0323	3.5	.0015
-1.1	.0641	1.3	.0261	3.7	.0011
- .9	.0784	1.5	.0209	3.9	.0008
- .7	.0868	1.7	.0166	4.1	.0006
- .5	.0894	1.9	.0130	4.3	.0005
- .3	.0873	2.1	.0101	4.5	.0003
- .1	.0817	2.3	.0078	4.7	.0002
.1	.0740	2.5	.0060	4.9	.0002
.3	.0652	2.7	.0046	Above 5.0	.0005

The various derived statistics herewith given have been computed from the exact distribution or from Salvosa's (1930) tabulation in .01 σ intervals so that an exact agreement with computations based upon Table VII A is not to be expected. We assume a sample of size 100.

$$M = .00;\ Mo = -.50;\ P_{.07} = -1.233;$$

$$P_{.50} = -.164;\ P_{.93} = 1.621;\ \frac{i_{.07}}{f_{.07}} = 3.854;$$

$$\frac{i_{.50}}{f_{.50}} = 2.383;\ \frac{i_{.93}}{f_{.93}} = 11.033;\ Pv = 2.854$$

$$V = 1.00;\ \mu_4 = 4.50;\ V_{pv} = .0849;\ V_v = .020;$$

$$As = .358;\quad Sk = .1254;\quad \mu_3 = 1.00;\quad \beta_1 = 1.00$$

$$V_{As} = .02506;\quad V_{Sk} = .00428;\quad V_{\mu_3} = .3600;\quad V_{\beta_1} = .6750$$

Critical Ratios

$$\frac{As}{\sigma_{As}} = 2.26;\quad \frac{Sk}{\sigma_{Sk}} = 1.95;\quad \frac{\mu_3}{\sigma_{\mu_3}} = 1.67;\quad \frac{\beta_1}{\sigma_{\beta_1}} = 1.22$$

The decreasing size of these four critical ratios is evidence of a decreasing efficiency in As, Sk, μ_3 and β_1 as measures of asymmetry. Also the less complicated measures As and μ_3 are more efficient than the abstract measures Sk and β_1. The efficiency of measures changes with change in form of distribution (See Fisher, 1921) *but it seems safe to believe that for a wide class of skewed distributions As is more efficient than μ_3 as a measure of asymmetry.* It is thus recommended for use in a test of asymmetry, but of course it is not serviceable for further computational work, for its algebraic combinatorial possibilities are very limited.

Comparison of mean and mode: As to how skewed a distribution should be to warrant the computation of mode rather than, or in addition to, the mean depends upon *(a)* which is the more reliable and *(b)* whether the distinction between them is possible and important for the issue in hand. The matter of reliability is accurately handled by comparing the standard error of the mode with that of the mean. Since there are several ways of computing the mode, both it and its standard error will depend upon the computational method followed. Pearson's method is to fit a curve with few constants (coefficients and exponents) to the entire distribution and then determine the mode and its standard error from this fitted curve. The merit of this procedure for distributions which are found by test (a χ^2 goodness-of-fit test of advanced statistics) well to fit a curve which is defined by just a few constants is granted, but the writer questions its general validity for psychological, educational, economic, and social data. In these fields the units of measurement are seldom natural, intrinsic, or inviolable. They are kept if found useful and discarded if not. Furthermore, there is generally no guarantee that the units which one happens to use retain a similar or constant merit throughout low, intermediate, and high ranges of values. We commonly, therefore, desire a mode which is not a function of the entire data, but only of that portion residing in the neighborhood of the mode. A computational method yielding such a mode, together with its standard error, is given later in this section. That this computational procedure is simpler than Pearson's is not the chief argument in its favor, which is that it yields a measure of central tendency which is important in itself, even though the entire distribution may not be defined and even though the significance of the units of

measurement change gradually when passing from low to high values. The standard error of the mode as thus computed may be less or greater than that of a Pearson mode, depending upon the nature of the distribution.

Maximum likelihood: A certain class of statistics is known as "maximum likelihood" measures. Think of a sample as derivable from any of a number of parent populations, all of the same type. The probability that the observed distribution is a sample from one of these is greater than that it is from any other one, so this particular parent may be called the maximum likelihood population. Its definition, i.e., the numerical values of the constants, or parameters, entering into its equation, is obtainable from the observed distribution. *A parameter so obtained is a maximum likelihood statistic.*

A unimodal population will yield sample distributions, and the observed mode of a sample is more likely to have arisen if the sample is one drawn from a parent with this same mode than if drawn from any other parent. Thus the sample mode, defining with maximum likelihood the parent mode, is a maximum likelihood statistic. This is true when the only assumption about the parent is that it has a mode.

The matter of maximum likelihood can be here merely touched upon, but this should suffice to inform the beginning student that the modal value of any statistic has certain important probability properties that are not possessed by any other value of the statistic.

That the mode is a natural point of reference is suggested by the simplicity of form that certain curves of the Pearson type take when the mode is made the origin. The writer has shown (1940) that certain invariant properties exist for modal age-grade norms that do not exist for other norms. Its general and genuine merit is

only clouded by the greater mathematical difficulty of manipulating maximum likelihood rather than mean statistics. For the reasons given it would seem that the distinction between the mode and the mean is an important one to make when it can be made with certainty and when an issue of central tendency is involved. We cannot relate this matter to the degree of skewness of a distribution, for a minute skewness is determinable if the sample is large enough. The statistical establishment of the mode as different from the mean is involved whether the mode is calculated by a fitted-curve method (in which case facilitating tables are given in Pearson's *Tables for Statisticians and Biometricians*) or by the simpler method here given. It would seem in general to be sufficient in the case of unimodal distributions to anticipate a useful distinction between mean and mode when asymmetry, as measured by *As* or μ_3, is established.

The computation of the mode: In Chapter IV an exercise was given in grouping so that the mode shown in a graphic presentation would have a known dependability. We here will give an algebraic method which is an extension of this practice. With fine grouping many modes characteristically appear in the sample and the crude mode has little to commend it over nieghboring values. With coarse grouping the possible sample crude mode values are few and far apart, so a close agreement with the population mode cannot be expected. The moving-average method of computing the mode is intended to remedy the defects of both fine and coarse grouping. Its process, merits, and defects can be illustrated by the data of Table VII B.

Certain of these lengths occur several times, but none should be called the crude mode, because none is outstanding. We first construct a frequency distribution, using a fine interval, so

TABLE VII B

HEAD LENGTHS, IN MM., OF 50 SIXTEEN-YEAR-OLD DELINQUENT BOYS

204	204	184	195	184
183	192	185	193	188
186	189	187	192	189
190	201	191	190	194
181	186	206	179	191
188	178	187	182	189
180	187	185	187	203
182	197	190	194	183
200	193	196	184	186
190	197	187	190	194

small that a computation error of one-half this interval, since the moving-average mode must be a class index or in the case of equal frequencies in neighboring classes a class boundary, is immaterial in the light of the problem and the size of the sample. We cannot choose an interval less than one millimeter, but this seems sufficiently small. We obtain the first two columns of Table VII C from Table VII B.

TABLE VII C

FREQUENCY TABLE OF LENGTHS, IN MM., OF HEADS OF 50 SIXTEEN-YEAR-OLD DELINQUENT BOYS, ALSO FREQUENCIES FOR DIFFERENT 2, 3, 4, and 5 MM. GROUPINGS

HEAD LENGTHS IN MM.	FREQUENCIES FOR VARIOUS CLASS INTERVALS					
	1 mm.	2 mm.	3 mm.	4 mm.	5 mm.	STANDARD 4 mm.
177.5						2
178	1					
179	1					
180	1					
181	1					
181.5						6
182	2					
183	2					
184	3		7		12	

TABLE VII C (CONTINUED)

HEAD LENGTHS IN MM.	FREQUENCY FOR VARIOUS CLASS INTERVALS 1 mm.	2 mm.	3 mm.	4 mm.	5 mm.	STANDARD 4 mm.
		5		10		
185	2		8		15	
185.5		5		13		13
186	3		10		15	
		8		12		
187	5		10		15	
		7		13		
188	2		10		18	
		5		15		
189	3		10		17	
189.5		8		12		12
190	5		10		14	
		7		12		
191	2		9		14	
		4		11		
192	2		6		14	
		4		9		
193	2		7		10	
193.5		5				8
194	3		6			
		4				
195	1					
196	1					
197	2					
197.5						3
198						
199						
200	1					
201	1					
201.5						3
202						
203	1					
204	2					
205						
205.5						3
206	1					
	50					50

The term, moving average, suggests the method of computation. For example, for the interval 5 mm. the first three values are obtained thus:

12 is the sum of 2, 2, 3, 2, 3
15 is the sum of 2, 3, 2, 3, 5
15 is the sum of 3, 2, 3, 5, 2, etc.

The first 15 may be gotten by dropping the 2 and adding the 5 to the value 12, the second 15 by dropping the 2 and adding the 2 to the first 15, etc.

With the 1 mm. interval there is no mode without a near competing mode, so a solution is not reached. Nor is there a solution with the 2 mm. grouping. With the 3 mm. interval a single mode is found and its value is 188.0. According to the usual procedure this would be taken as the answer. The coarser groupings are given to be illustrative of the uncertainty of the solution. No unambiguous mode is revealed with the 4 mm. grouping, but the value 188.0 appears with the 5 mm. interval. Some follow the practice of taking moving averages of moving averages, with doubtful improvement in outcome. The standard error of the mode computed by any of the moving-average methods is unknown.

We will compute the mode by determining the maximum point of the best fit parabola to the frequencies in the four neighboring classes whose sum of frequencies is greatest. The data of Table VII E shows that this parabolically smoothed mode is a very serviceable approximation to the Pearson fitted curve mode. As one would scarcely consider using the mode as a measure of central tendency in the case of a flat-topped distribution, the failure of formula [7:24] in such cases is not important.

The size of the interval employed is of prime importance. We will determine the number of

classes and the class boundaries according to the rules already given in Chapter IV for graphic portrayal. These have, in general, been found to be meritorious for the purpose of computing the mode. We cannot say that the interval given by these rules is the best, for what is best depends upon the nature of the true distribution for the phenomena in question. Investigation, for the normal distribution, of the optimal size of interval to use in determining the parabolically smoothed mode so as to minimize the standard error of the discrepancy between the frequency at the mode as given by the parabola and the true modal frequency has resulted in the values of Table VII D. The computational procedure was such that the results are only approximate, though the error in the intervals given is certainly less than $.1\sigma$.

TABLE VII D

OPTIMAL INTERVAL, IN TERMS OF THE POPULATION σ, FOR THE COMPUTATION OF THE PARABOLICALLY SMOOTHED MODE, FOR SAMPLES OF SIZE N DRAWN FROM A NORMAL POPULATION

N	22	30	60	125	200	400	1300
Size of interval	1.1	1.0	.9	.8	.7	.6	.5

A comparison with Table XIII H shows that these intervals are considerably larger than there given. Obviously the more leptokurtic the curve the smaller the optimal interval when expressed in terms of the standard deviation, and as typically the merit of the mode is associated with skewed leptokurtic data, the smaller (though still wide) intervals of Table XIII H are to be preferred to those of Table VII D.

Having found a practical solution for this troublesome matter of size of interval, we will now give the necessary formulas and compute:

Mo = the parabolically smoothed mode

σ_{Mo} = the standard error of this mode

f_{Mo} = the modal frequency

and σ_{Mo} = the standard error of this frequency.

From the general nature of most non-cuspidal unimodal frequency distributions of continuous variables, it is inferred that a second-degree parabola can be made to fit closely to these curves in the neighborhood of their modes. We will accordingly determine the parabola that best fits the four points closest to the mode to be determined. Let us use an arbitrary scale, called a ξ scale, in terms of which the values of the class indexes of the four classes in question are -1.5, $-.5$, $.5$, and 1.5 respectively, and let the frequencies of these classes be represented by f_1, f_2, f_3, and f_4. One unit upon this arbitrary scale in ξ is equal to i units in the original data, and the origin of the ξ scale is the boundary between the second and third classes. The algebraic relation between X and ξ is

$$X = \text{Arbitrary origin} + i\,\xi \quad . \quad . \quad . \quad . \quad . \qquad [7:20]$$

If the mode is determined in ξ units, then its value in X units is given by

$$Mo_x = \text{Arbitrary origin} + i\,Mo_\xi \quad . \quad . \quad . \quad . \qquad [7:21]$$

The equation of the parabola which best fits (in the least squares sense) the four points $(-1.5, f_1)$, $(-.5, f_2)$, $(.5, f_3)$, $(1.5, f_4)$ is

$$f = \frac{1}{16}(-f_1 + 9f_2 + 9f_3 - f_4) + (-.3f_1 - .1\,f_2 + .1\,f_3 + .3f_4)\xi + \frac{1}{4}(f_1 - f_2 - f_3 + f_4)\xi^2 \qquad [7:22]$$

The equation of the parabola which best preserves the areas, in the least squares sense, in the four intervals is

$$f = \frac{1}{12}(-f_1 + 7f_2 + 7f_3 - f_4) + (-.3f_1 - .1\ f_2 + .1\ f_3 + .3f_4)\xi + \frac{1}{4}(f_1 - f_2 - f_3 + f_4)\xi^2 \qquad [7:23]$$

which, it will be noted, differs from [7:22] only in the constant term. It thus has the same mode as [7:22].

The mode, or value of ξ for which the frequency is a maximum, is

$$Mo_\xi = \frac{-.6f_1 - .2f_2 + .2f_3 + .6f_4}{-f_1 + f_2 + f_3 - f_4} \quad \text{The mode in } \xi \text{ units} \qquad [7:24]$$

Accordingly, substituting this value in [7:21] yields the answer which we will call *the parabolically smoothed mode*. **If the denominator of** [7:24] is negative the sample has an anti-mode in this region, so we have evidence of bimodality and should consider the method inapplicable.

Substituting Mo_ξ for ξ in [7:23] yields the modal frequency per i interval

$$f_{mo} \text{ per } i \text{ interval} = \frac{-f_1 + 7f_2 + 7f_3 - f_4}{12} + \frac{-f_1 + f_2 + f_3 - f_4}{4}(Mo_\xi)^2 \qquad [7:25]$$

Of course

$$\frac{f_{mo} \text{ per } i \text{ interval}}{i} = f_{mo} \text{ per unit (in } X) \text{ interval} \qquad [7:26]$$

The variance error of Mo_ξ is approximately given by [7:27] .

$$V_{Mo_\xi} = \frac{f_1 + f_2 + f_3 + f_4}{(-f_1 + f_2 + f_3 - f_4)^2} (.2 + Mo_\xi^2) \quad \text{Variance error of } Mo_\xi\text{*} \quad [7:27]$$

$$V_{Mo} = i^2 V_{Mo_\xi} \quad \text{Variance error of parabolically smoothed mode. . . . .} [7:28]$$

* Calculation of the standard error of the parabolically smoothed mode

We have

$$Mo_\xi = \frac{-.6f_1 - .2f_2 + .2f_3 + .6f_4}{-f_1 + f_2 + f_3 - f_4} = \frac{Num}{Den}$$

Taking logarithmic differentials (see Chapter XIII, Section 8) we have

$$\frac{d\,Mo_\xi}{Mo_\xi} = \frac{d\,Num}{Num} - \frac{d\,Den}{Den}$$

Squaring, summing, and dividing by the number of samples summed, we have

$$\frac{V_{Mo_\xi}}{Mo_\xi^2} = \frac{V_{Num}}{Num^2} + \frac{V_{Den}}{Den^2} - \frac{2c_{Num,Den}}{Num\ Den}$$

We let $\frac{f_1}{N} = p_1$ and $q_1 = 1-p_1$. Similar definitions hold for p's and q's with other subscripts. We also let

$$p = (p_1 + p_2 + p_3 + p_4)/4$$

Utilizing relationships of the types

$V_{f_1} = N p_1 q_1$ and $\sigma_{f_1} \sigma_{f_2} r_{f_1 f_2} = - N p_1 p_2$

$$V_{Num} = N(.36p_1q_1 + .04p_2q_2 + .04p_3q_3 + .36p_4q_4 - .24p_1p_2 + .24p_1p_3 + .72p_1p_4 + .08p_2p_3 + .24p_2p_4 - .24p_3p_4) \doteq .8Np$$

As a check upon the excellence of the mode as given by [7:24] and [7:21] we may compare it as thus computed with the value as determined by more refined methods. Volume one of *Biometrika* was scanned to find illustrations of data to which Pearson curves had been fitted and the mode determined from the fitted curves. Articles by Powys (1901), Macdonnell (1901), Pearson (1902, Sys.), and Fawcett (1902) serve excellently. The data by Powys are particularly valuable because of the large size of samples. Testing the parabolically smoothed mode against the mode from the fitted curve for these large samples is a severe test so far as systematic differences are concerned. Examination of the results as shown in Table VII E reveals no systematic error, and it further shows that the chance error is adequately represented by the formula given, [7:28], for the variance error of the parabolically smoothed mode.

FOOTNOTE CONTINUED:

$$V_{\text{Den}} = N(p_1q_1+p_2q_2+p_3q_3+p_4q_4+2p_1p_2+2p_1p_3-2p_1p_4$$

$$-2p_2p_3+2p_2p_4+2p_3p_4)$$

$$\doteq 4Np$$

$$c_{\text{Num, Den}} = N(.6p_1q_1-.2p_2q_2+.2p_3q_3-.6p_4q_4+.4p_1q_2+$$

$$.8p_1q_3+0+0-.8p_2p_4-.4p_3p_4)$$

$$\doteq 0$$

Collecting terms

$$V_{Mo_\xi} \doteq \frac{4Np}{Den^2}\,(.2+Mo_\xi^2) = \frac{(f_1+f_2+f_3+f_4)}{(-f_1+f_2+f_3-f_4)^2}(.2+Mo_\xi^2)$$

TABLE VII E

COMPARISON OF FITTED CURVE AND PARABOLICALLY SMOOTHED MODES

FOOT-NOTE	SIZE OF SAMPLE	FITTED CURVE METHOD			PARABOLICALLY SMOOTHED MODES				
								No. OF CLASSES	
		PEAR-SON TYPE	MODE	FREQUEN-CY AT MODE	MODE	σ_{Mo}	FREQUEN-CY AT MODE	AVAIL-ABLE	PRE-FERRED
(a)	5542	IV	66.95	168.6	66.98	.07	172.9	23	30
(b)	12747	IV	67.40	161.5	67.32	.02	164.2	27	44
(c)	15117	IV	67.16	156.4*	67.21	.014	164.5	25	45
(d)	9003	IV	66.96	159.2	67.01	.02	161.5	24	39
(e)	4407	IV	66.82	154.4	66.97	.11	155.5	22	30
(f)	2862	IV	66.39	151.9	66.27	.16	153.4	21	25
(g)	3000	IV	15.02	239.4	15.05	.07	232.5	35	30
(h)	3000	IV	65.48	480.6	65.66	.17	469.8	22	30
(i)	2000	I	.653	171.#	.647	.013	167.1	15	24
(j)	500	III	11.32	98.5#	10.05*	2.4≠	99.0	13	17
(k)	181	I	131.5	14.8#	130.1	2.7	14.1	13	13

* Visual inspection of graph indicates a poor fit at this point.

\# Read from graph.

≠ f_1 is greater than f_2 or f_3 or f_4 so a poor fit to curve entire is to be expected. This is reflected in size of σ_{Mo}.

(a) Powys: Stature, in inches, of New South Wales male criminals, Ages 20-25.
(b) Powys: Stature, in inches, of N.S.W. male criminals, ages 25-30.
(c) Powys: Stature, in inches of N.S.W. male criminals, ages 30-40.
(d) " " " " " " " " " 40-50.
(e) " " " " " " " " " 50-60.
(f) " " " " " " " " " 60- .
(g) Macdonnell: Head breadth in Cm. of Criminals (Scotland Yard).
(h) " Stature, in inches, " " " "
(i) Pearson: Fecundity of broodmares.
(j) " Thiele's observations from game of "patience."
(k) Fawcett: Breadth of female skull in mm.

A general formula for the variance error of the parabolically smoothed modal frequency is involved. It is simple for the special case in which the true value of $Mo_{\xi} = 0$. We then have

$$V(f_{Mo} \text{ per } i \text{ interval}) = \frac{1}{144N}\,[f_1(N-f_1) + 49f_2(N-f_2) + 49f_3(N-f_3) + f_4(N-f_4) + 14f_1f_2 + 14f_1f_3 - 2f_1f_4 - 98f_2f_3 + 14f_2f_4 + 14f_3f_4] \quad \text{Variance of } f_{Mo} \text{ when } Mo_{\xi} = 0. \qquad [7:29]$$

When Mo_{ξ} does not equal zero all we can assert is that the variance error of (f_{Mo} per i interval) is greater than the value given by [7:29].

SECTION 4. THE GEOMETRIC MEAN

The use of the geometric mean as an average might be suggested by discovery that, for the data in question, it was more reliable than the other common averages. However, the determination of this may be a very complicated undertaking. In ordinary practice the logic of the situation rather than an objective measure of reliability has been the consideration which has led to the use of the geometric mean. Though in general we believe this has been sound practice, the statistician would be more satisfied if the use of the geometric mean were fortified by demonstrable evidence of reliability.

Data for which the geometric mean seems appropriate have the following characteristics: *(a)* The raw measures are all positive and are deviations from an indubitable and meaningful zero point. *(b)* Thinking is facilitated when relative, not absolute, amounts are kept in mind. *(c)* It follows as a consequence that certain tests of consistency (e.g., the time and quantity reversal tests applied to index numbers) are

frequently satisfied by the geometric mean and not by alternative averages. In the following discussion the reader should think of the raw X scores as fulfilling conditions *(a)* and *(b)*.

If $X_a, X_b, X_c, \ldots X_n$ are the measures in a series, and if πX stands for the product of them all, we have, by definition,

$$G.M. = \sqrt[n]{\pi X} \quad \ldots\ldots\ldots\ldots \quad [7:30]$$

Taking logarithms we have

$$\log(G.M.) = \frac{\log X_a + \log X_b + \ldots + \log X_n}{N} = \frac{\Sigma \log X}{N}$$

$$= \frac{\Sigma Y}{N} = M_y$$

where Y is the logarithm of X. Thus corresponding to a geometric mean of the X's is an arithmetic mean of the Y's and all the error and probability properties attaching to the arithmetic mean can, by utilizing the one-to-one relationship that exists, be attached to the geometric mean. The probability that $\widetilde{G.M.}$ (the true $G.M.$) exceeds the obtained $G.M.$ by an amount Δ is exactly the probability that $\tilde{M}_y$ exceeds M_y by an amount δ, the relationships between quantities being

$$\text{Log } \widetilde{G.M.} = \tilde{M}_y;\ \log G.M. = M_y;\ \Delta = \widetilde{G.M.} - G.M.;$$

$$\delta = \tilde{M}_y - M_y$$

The form of distribution of the Y measures and the variance error of M_y $(= \frac{Vy}{N-1})$ are as simple to determine as the form of distribution and mean of any raw measures. This relationship

between measures and their logarithms suggests the following practice:

When the geometric mean is used, interpret relationships between measures in relative terms, but judge of the confidence to be placed in the geometric mean via the confidence to be placed in its logarithm.

Confidence as here used is that due to the nature and size of the sample. In many, perhaps most, situations wherein the geometric mean is appropriate a sample does not exist. If the population of a city in 1930 is 152760 and in 1940 it is 186210, we do not say that we have a sample of two. We have two observations in a time series and these observations at different moments of time are not assumed to be two observations of the same thing drawn from an infinite population. We may ask, "What is the average yearly rate of growth?" The ten-year growth is 33450. We do not divide this by ten and call the yearly growth 3345, but assume a constant yearly rate of growth.

Population in 1930	=	152760
Assume in 1931 it	=	$152760(1+r)$
Assume in 1932 it	=	$152760(1+r)^2$
etc.		
Assume in 1940 it	=	$152760(1+r)^{10} = 186210$

Taking logarithms and solving, $r = .0200$.

We find an average increase of 2 per cent a year.

Had we the following data: Mean vocabulary of 100 I.Q. sixteen-year-olds 15276 words, and of 110 I.Q. sixteen-year-olds 18621 words, we would not have dealt with rates but have said that there was an average increase of 334.5 words per unit increase in I.Q. We note that which procedure is appropriate has depended upon our

general knowledge of the situations which are considerations outside the numerical values of the measures themselves. We may believe this approach justified when dealing with statistical series which are not similar measures drawn from an infinite population. This situation is very common when dealing with temporal and geographic data.

The time reversal test is of proven importance in connection with economic indexes. One would anticipate its importance in connection with fatigue, learning, growth, and sundry sociological phenomena of a temporal nature, provided measurements from a true zero point are available. In economics the test *requires that if a price (or quantity) index based upon a number of commodities shows a certain relationship between a first and a second date, the reciprocal relationship is to be shown between the second and the first date.* Consider the following data covering a stable, 1913, and an inflated, 1918, period.

TABLE VII F

PRICES AND PRICE RATIOS OF CERTAIN COMMODITIES, 1913, 1918

	Prices		Ratios	
	1913	1918	1913 on 1918	1918 on 1913
Bacon	.1236	.2612	.47320	2.11327
Pork	.1486	.2495	.59559	1.67900
Lard	.1101	.2603	.42297	2.36421
Arithmetic price index	.12743	.25700	.49725* .49584	2.05216*
Geometric price index	.12646	.25694	.49216# .49215	2.03188#

* These values are the arithmetic means of preceding ratios.

\# These values are the geometric means of preceding ratios.

The ratio of the price of bacon in 1913 to that in 1918, .47320, is the reciprocal of the ratio of the 1918 price to that of 1913, 2.11327, so the time reversal test holds for prices of single commodities. The arithmetic average of the 1913 ratios, .49725, is not the reciprocal of the arithmetic average of the 1918 ratios, 2.05216, so the time reversal test does not hold for these mean ratios. Also the value of the mean 1913 on 1918 ratios, .49725, does not equal the ratio of the means, .49584.

In the case of the geometric mean of the price ratios, the time reversal test does hold for .49216 is the reciprocal of 2.03188. Also the ratio of the geometric price indexes, .49216, is identical with the geometric mean of the separate price ratios. It is left to the student as an exercise to prove that this is necessarily so.

In this illustrative example the three commodities have been weighted equally (each weighted 1), but clearly if weighted unequally all that is necessary is that each commodity be taken as many times as the weight, and then, proceeding as here shown, we again find the time reversal test holding for the geometric means. There are other price indexes than the geometric mean for which the time reversal test holds, but none which are more simple algebraically.

The ideal distribution of the cases constituting a sample in which to use the geometric mean is one in which the logarithms of the measures are distributed normally. The data of Table VII G yield such a distribution.

Because of the extreme skewness of these data it has been necessary to report frequencies for unequally spaced intervals. A plot of these data as a frequency polygon will be informative in strikingly revealing the form of distribution for which the geometric mean is most appropriate. If logarithms of these X measures to the base e

TABLE VII G

DISTRIBUTION OF 1000 MEASURES THE LOGARITHMS OF WHICH YIELD A NORMAL DISTRIBUTION

X	ORDINATE z	INTERVAL i	FREQUENCY PER INTERVAL = 1000 z i
.167	.0008	.166667	+
.333	.0099	.166667	2
.500	.0212	.166667	4
.667	.0331	.166667	6
1.0	.0539	.5	27
1.5	.0745	.5	37
2.0	.0848	.5	42
2.5	.0886 (e=mode)	.5	44
3.0	.0884	.5	44
3.5	.0861	.5	43
4.0	.0825	.5	41
4.5	.0782	.5	39
5.0	.0738	.5	37
5.5	.0693	.5	35
6.0	.0653	.5	33
6.5	.0608	.5	30
7.0	.0568 (e^2=Mdn)	.5	28
7.5	.0531	.5	27
8.0	.0496	.5	25
8.5	.0464	.5	23
9.0	.0434	.5	22
9.5	.0406	.5	20
10.5	.0357	1.5	54
12.0	.0295	1.5	44
13.5	.0246	1.5	37
15.0	.0207	1.5	31
16.5	.0175	1.5	26
18.0	.0149	1.5	22
20.0	.01218	2.5	31

TABLE VII G (CONTINUED)

X	ORDINATE z	INTERVAL i	FREQUENCY PER INTERVAL = 1000 z i
22.5	.00958	2.5	24
25.0	.00766	2.5	19
27.5	.00619	2.5	15
30.0	.00506	2.5	13
32.5	.00409	2.5	10
35.0	.00339	2.5	9
38.0	.00274	3.5	10
42.0	.00210	4.5	10
47.0	.00153	5.5	8
53.0	.00108	6.5	7
60.0	.00074	7.5	6
68.0	.00050	8.5	4
77.0	.00033	9.5	3
87.0	.000219	10.5	2
98.0	.000144	11.5	2
112.0	.000088	16.5	2
133.0	.000046	25.5	1
171.0	.000017	50.5	1
249.0	.000003	101.5	+
450.0	.000000+	300.5	+
			1000

are found and a distribution made, it will be normal, with a mean of 2.0 and a variance of 1.0, and these statistics will have small standard errors. For the X measures the mean and variance would be, of course, very poor statistics.

SECTION 5. THE HARMONIC MEAN

By definition the harmonic mean, HM, of a series of measures, X, is the reciprocal of the

mean of the reciprocals, thus,

$$\frac{1}{HM} = \frac{1}{N}\Sigma\frac{1}{X} \qquad \text{Harmonic Mean} \qquad [7:31]$$

All X measures must be positive. Let $Y = 1/X$. To compute the HM we compute the Y values, obtain their mean, and take the reciprocal. Having a distribution of Y values we can obtain M_y, V_y, and V_{M_y} by procedures already given. Thus, to *obtain a measure of confidence in the HM, we first secure a measure of the reliability of M_y. For any fiducial limits in HM that we may be interested in, there are corresponding limits in M_y and we judge the trustworthiness of M_y via methods already given for testing the significance of the arithmetic mean.*

The HM is serviceable in connection with "work limit" measures. In the phraseology of psychology a test calling for a fixed amount of work, and scored on the basis of time required to accomplish it, is called a "work limit" test. It stands in distinction to a "time limit" test which is one having a fixed time limit and scored on the basis of amount done. Let us consider the case of individuals A, B, and C, able to accomplish 30, 40, and 50 units of work, respectively, in one hour. If tested by means of a half-hour time-limit test, their amount-done scores will be 15, 20, and 25. If tested by means of a 60-item work-limit test, their time-required scores will be 120, 90, and 72 minutes respectively,—showing quite different relationships than do the time-limit test scores. The time-limit data yield an average score of 20 units in 30 minutes, which is at the rate of one unit in 1.5 minutes. We desire to get the same information from the work limit scores, but their mean, 94, equivalent to a rate of one unit done in 1.5667

minutes, does not give it. For the *HM* we have

$$\frac{1}{HM} = \frac{1}{3}\left(\frac{1}{120} + \frac{1}{90} + \frac{1}{72}\right) = \frac{1}{90}$$

Thus the *HM* time required to perform 60 units of work is 90 minutes, which is at the rate of one unit in 1.5 minutes,—checking with the time-limit test results.

To determine which mean to employ the student must decide for his particular situation which mean is the more immediately interpretable,—that whose base is constant time (the time-limit test score), or whose base is constant product (the work-limit test score). Ordinarily, the meaning in the reader's mind of a unit of time is both precise and comparable from one situation to another, but the meaning of a unit of work, even when precise, as e.g. , might be one ton of coal, one automobile, one ton-mile of transportation, is not comparable from situation to situation. Thus the *HM* is suggested as the appropriate average to use in a work-limit test in psychology or in industry, where a common procedure is to record the time it takes a workman to perform a unit of work.

Another obvious consideration is to select the average with the smaller standard error. If the *Y* measures are more nearly normally distributed than the *X* measures, the *HM* will generally be more trustworthy than the *M*.

It was noted that the *GM* price index yields consistent results as changes in the basal year are made, and that in this respect there is a systematic error in the arithmetic mean index. There is a systematic error in the *HM* price index, but it is of the opposite sort to that in the *M* price index. As a result, the average (preferably geometric) of an *M* and an *HM* price index is unbiased.

PROBLEMS

Problem 1. Prove that the G.M. of ratios of measures in two paired series is equal to ratio of the G.M.'s of the two series. Suggested notation follows:

Let p_{1a} = price of commodity a at date 1, and similarly for $p_{1b} \cdots p_{1n}$, p_{2a}, $p_{2b} \cdots p_{2n}$, p_{3a}, $p_{3b} \cdots p_{3n}$, etc. Then p_{1a}/p_{3a} = a price ratio of commodity a at date 1 to 2. Any average of these for several commodities, which we will call I_{12} is a price index. Problem: to find such an average that $\dfrac{I_{12}}{I_{32}} = I_{13}$. This asserts, for example, that if the price indexes using 1920 as the basal year ($I_{1910,1920}$, $I_{1911,1920}, \cdots I_{1930,1920}$) are available, the price indexes for any other basal year, say 1930, are available

for $I_{1910,1930} = \dfrac{I_{1910,1920}}{I_{1930,1920}}$, $I_{1911,1930} = \dfrac{I_{1911,1920}}{I_{1930,1920}}$, etc.

Prove that the simple G.M. of price ratios is such an average. Also prove that the arithmetic mean of price ratios is not.

CHAPTER VIII

THE NORMAL DISTRIBUTION

SECTION 1. THE NORMAL DISTRIBUTION AS DESCRIPTIVE OF CHANCE DISTRIBUTIONS AND DISTRIBUTIONS OF ERRORS

Phenomena of great diversity have suggested the normal distribution to the mind of man. It inevitably came to the attention of analytical gamesters interested in the probability of events when coins were tossed, dice thrown, cards drawn from a deck, and roulette wheels spun. An important relationship inherent in all of these cases is that, starting with distributions of raw data which are rectangular, or point rectangular, one approaches normal distributions of statistics which are aggregates or averages. If an unbiased die is thrown many, say N, times we approach the following distributions:

Number of pips	1	2	3	4	5	6
Number of occurrences	$\frac{N}{6}$	$\frac{N}{6}$	$\frac{N}{6}$	$\frac{N}{6}$	$\frac{N}{6}$	$\frac{N}{6}$

which we call a point rectangular distribution. This certainly has no semblance to a normal dis-

tribution. Consider a still simpler case: the number of heads showing when an unbiased coin is tossed N times the distribution will approach the following:

Number of heads	0	1
Number of occurrences	$\frac{N}{2}$	$\frac{N}{2}$

If we throw 20 coins at a time and compute for each throw the mean number of heads, we will approach the binomial distribution given by the successive terms in the expansion of $(.5 + .5)^{20}$ which is very nearly normal. Though the distribution of single items is two point rectangular, the distribution of means of such tends toward normality. The writer knows of no exception to the proposition that *the distribution of statistics which are means or aggregates of original measures, no matter their form, tends toward normality if the number of cases entering into the mean or aggregate is large.*

The normal distribution is a limiting form and the number of paths whereby to approach it are infinite. A. DeMoivre (1733) is to be credited with being the first mathematician to derive the normal curve and compute probabilities based thereon.

At an early date astronomers found that the distribution of their angular observations of a star, which could only be considered to be errors of observation, approached normality as the number of observations increased. We have here a phenomenon that is normally distributed, though it is not an average or aggregate. *In general we may anticipate that when a fixed point is aimed at, the distribution of errors resulting will either be normally distributed or that some simple transformation of these measures will be*

so distributed. To cite two cases involving transformed values: If lifted weights are compared with a standard weight the distribution of errors, expressed in grams, will not be as closely normal as will the distribution of errors if the logarithms of weights have been employed. Again, if teachers attempt to classify pupils according to mental age, attempting, say, to put into a single group all those of a certain mentality, the distribution of their errors expressed in terms of mental age will not be as closely normal as will the distribution expressed in sensed difference units. The writer offers the hypothesis that all errors of observation are normally distributed when expressed in units which are intrinsically appropriate to the sense organ or faculty making the observation.

SECTION 2. THE NORMAL DISTRIBUTION AS DESCRIPTIVE OF BIOLOGICAL AND SOCIAL PHENOMENA

Abraham DeMoivre, the discoverer in 1733 of the normal curve as a limiting form to the binomial $(a+b)^n$, also conceived of it in the broadest manner for deviations from it were no less than fluctuations from the original design of the Deity. It was not only descriptive of the universe of physics, but also of biology, sociology, philosophy, and religion, DeMoivre's successors, scarcely held this breadth of view, though Laplace later in the same century and Gauss, beginning with his *Theoria motus corporum coelestium* in 1809, both greatly extended the concept in connection with mathematics and the physical universe. Still later the astronomer Quetelet turned his attention to human affairs, writing, in 1846, *Lettres . . . sur la théorie des probabilités, appliquée aux sciences morales et politique*, and in 1871 upon *Anthropométrie ou Mesure des Différentes Facultés de l'Homme*.

Herewith is a striking table prepared by Quetelet:

TABLE VIII A

QUETELET'S TABLE OF HEIGHTS OF AMERICAN CIVIL WAR SOLDIERS

MESURES de HAUTEUR METRIQUE*	NOMBRE des recenses, pr difference de hauteur, de $0^{m}0255$.	PROPORTION de la hauteur de 1000 inscrits mesures. OBS.	CALCUL.	DIFFERENCES entre les valeurs observees et calculees
1,397 a 1,524 . .	31	1	2	-1
1,549	15	1	3	-2
1,575	50	2	9	-7
1,600	526	20	21	-1
1,626	1237	48	42	+6
1,651	1947	75	72	+3
1,676	3019	117	107	+10
1,702	3475	134	137	-3
1,727	4054	157	153	+4
1,753	3631	140	146	-6
1,778	3133	121	121	0
1,803	2075	80	86	-6
1,829	1485	57	53	+4
1,854	680	26	28	-2
1,880	343	13	13	0
1,905	118	5	5	0
1,930	42	2	2	0
1,956[and over] .	17	1	0	+1
	25878	1000	1000	-28 +28

* INTERNATIONALER STATISTISCHER CONGRESS IN BERLIN, t. II, page 748.

Could any student, if the first to make the tabulations shown in Table VIII A, fail to be struck with wonder at the evidence of a permeating design? Quetelet wrote of this and other data (1871):

"It is especially noteworthy that human height, though apparently having been developed in a fortuitous manner, is nevertheless subject to a very exact law; and this is not peculiar to height, but is observable also in connection with weight, strength and speed of man, and further in his intellectualy and moral qualities. We see one of the most wonderful laws of creation in this orderly diversity which is accomplished entirely without the intervention of the human will." This was both a return to the breadth of view of DeMoivre and an immediate stimulus to Francis Galton whose studies of hereditary genius became enriched with the concepts of normal distribution and correlation.

In Galton's *Natural Inheritance* (1889) is found a Table entitled "Data for Schemes of Distribution of Various Qualities and Faculties among the Persons Measured at the Anthropometric Laboratory of the International Exhibition of 1884." Galton used the median as the measure of central tendency and Q′, a smoothed estimate of the quartile deviation, as the measure of variability. Using these and the normal distribution he derived measures as shown in Table VIII B.

With the more refined present-day methods of testing goodness of fit the statistician could show that some of the traits of Table VIII B are not normally distributed, but nevertheless the deviation is not great and the major impression gotten by Galton, that the traits are normally distributed, was a great advance over that of his contemporaries who, except for a few, must have held either no view or a decidedly erroneous one.

The nature of the distribution of mental traits is, of course a function of *(a)* the method of selection of the group examined, *(b)* the units of measurement employed, and *(c)* the precision, or lack of chance factors, in the measure

TABLE VIII B

DEVIATIONS FROM THE MEDIAN IN TERMS OF Q′

Subject of Measurement	Age	Unit and Q′	Sex	*N*	Percentiles					
					5	10	30	70	90	95
		Inches								
Ht. standing without shoes	23-51	1.72	m	811	2.73	1.98	.81	.76	1.98	2.61
		1.62	f	770	2.71	2.10	.74	.80	1.91	2.46
Ht. sitting	23-51	.95	m	1013	2.52	1.89	.73	.73	1.79	2.31
		.82	f	775	2.55	1.95	.73	.85	2.07	2.55
Span of arms	23-51	2.07	m	811	2.36	1.83	.82	.72	1.79	2.36
		1.87	f	770	2.35	1.87	.69	.80	1.98	2.67
		Pds.								
Wt. in ordinary indoor clothes	23-26	10.00	m	520	2.20	1.80	.80	.70	2.20	2.90
		11.00	f	276	1.80	1.60	.70	.90	1.80	2.40
		Cu. in.								
Breathing capacity	23-26	24.50	m	212	2.32	1.68	.80	.68	2.32	2.84
		19.00	f	277	2.39	1.87	.73	.67	2.03	2.49

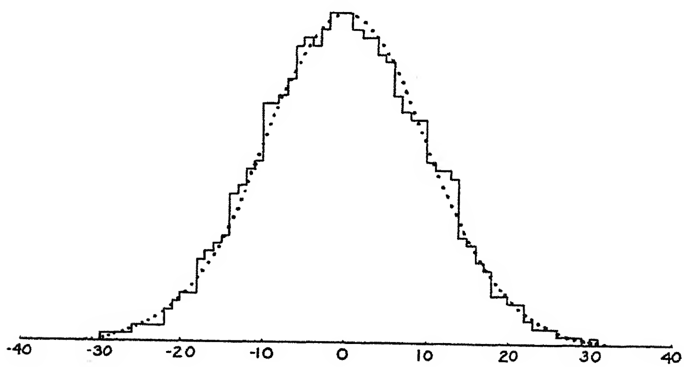
Strength of pull as archer with bow	23-26	PDS. 7.50	m	519	2.39	1.86	.80	.80	1.99	2.92
		5.22	f	276	1.92	1.06	.53	.53	1.46	1.86
Grip, strongest hand	23-26	7.75	m	519	2.32	1.81	.77	.77	1.93	2.45
		7.50	f	276	2.12	1.73	.66	.80	1.99	2.66
Swiftness of blow	23-26	FT. PER SEC. 2.37	m	516	2.06	1.68	.80	.80	1.77	2.31
		1.55	f	271	2.71	2.13	.84	.71	1.87	2.26
Keenness of vision distance of reading test type	23-26	INCHES 4.00	m	398	3.00	2.00	.75	.75	1.75	2.25
		5.22	f	433	2.66	2.28	.95	.57	1.33	1.52
Sums					43.11	33.12	13.65	13.34	33.96	43.82
Means multiplied by 1.015 to make $Q = 1$					2.44	1.87	.77	.75	1.92	2.47
Normal values when . . . . $Q = 1$					2.44	1.90	.78	.78	1.90	2.44

employed. If there are large chance errors we should anticipate that, like errors of observation, they would tend to be distributed normally. Thorndike and Bregman, (1924), after making adequate allowance for the normalizing influence of chance factors, concluded "that intellect in the ninth grade, if measured in truly equal units, is distributed approximately in [the normal manner]." This finding pertained to "word tests," "space tests," and a more general test involving "selective and relational thinking, and generalization and organization" separately, and also to a composite based upon nine mental test measures given to some 14,000 widely distributed ninth-grade pupils. Their composite distribution is shown in Chart VIII I.

CHART VIII I

COMPOSITE CURVE FOR THE NINTH GRADE

BASED UPON NINE SINGLE CURVES. THE BROKEN LINE INDICATES THE THEORETICAL NORMAL CURVE.



In all these cases of normality mentioned, and in innumerable other cases which could be cited, it is obvious that there has been a selective and environmental pressure operating. Geneticists have commonly secured point, or near point, distributions, but their plant and animal products are favored and not subject to the laws of survival of the plant and animal in the wild state. Many,—probably most and possibly all,—of the mutative point characters of the wild-type fruit fly are noncontributive to survival in the wild state. Institutional care of the feeble-minded fruit fly, the eyeless one, the wingless one, etc., permits survival. In the wild type such essential characters as wing-spread, weight, etc., as condition survival betray unimodal symmetrical or slightly skewed distributions. The experimental facts of genetics and medicine and the observable facts of stable plant, animal, and human life suggest that elementary causes (genes, types of infection, specific foods, vitamins, etc.) which, if they could be singly operative, would lead to point phenomena, do in fact so combine with innumerable other elementary or complex causes as to yield adjustment and survival traits which are of the unimodal symmetrical or slightly skewed types. A notable exception to this is sex, which characteristically yields a dromedary-back distribution, one hump for male and the other for female. Except for certain mental associates of sex and certain non-sex linked characteristics which are unimportant for survival, such as taste sensitivity, psychological investigation and experimentation have not as yet established the existence of mental traits of the dromedary type, though innumerable researches have been such as to permit the betrayal of bi- or multi-modality, if present. The characteristic distributions of sociology and economics are unimodal, as also are most, but not all, of those

of physiology and medicine. We cannot conclude that normality of distribution is universal in biological and social phenomena, though we may confidently expect to find distributions which do not differ greatly from it.

SECTION 3. THE NORMAL DISTRIBUTION AS RELATED TO ADVANCED STATISTICAL THEORY

Every experimental situation reduces to that of the significance of some difference. There are two aspects of significance, *(a)* the magnitude and *(b)* the trustworthiness of the difference. Depending upon the problem, the practical issue is *(a)* or *(b)* or an issue involving both. To interpret magnitude some form of parent population must be postulated, and in many situations this is the normal distribution. To judge of significance some law governing the errors of observation and of measurement must be postulated, and for cogent reasons this is taken to be a normal distribution of errors in large samples and the appropriate derivatives therefrom for small samples. The t-distribution, shown through the courtesy of Dr. Phillip J. Rulon, in Chart VII II, is appropriate for interpreting the

CHART VIII II

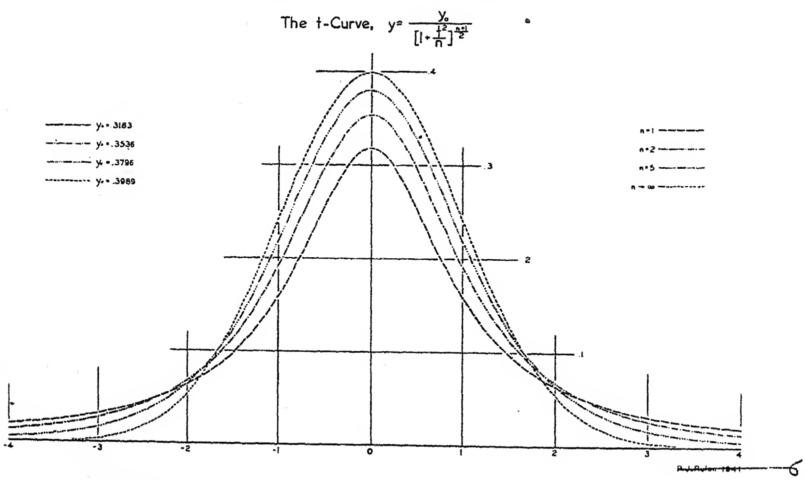
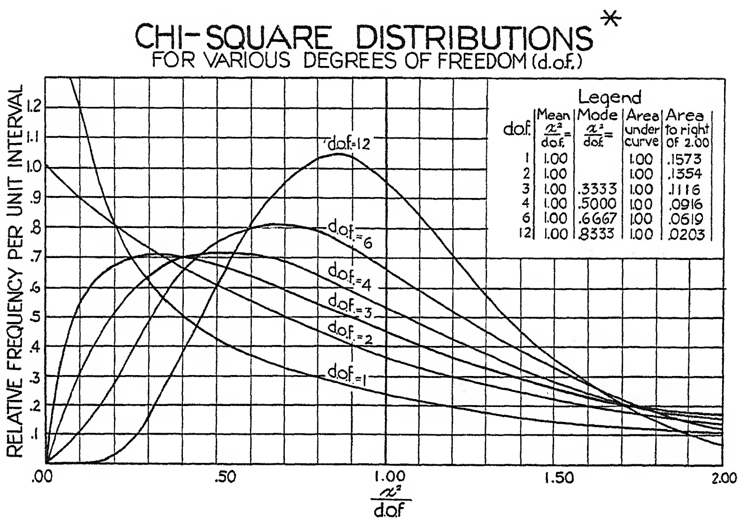


CHART VIII III



* Let x_1, x_2, . . x_j be uncorrelated, but associated, variables (i.e., r_{12}, etc. = 0), each having mean of zero, a standard deviation of one, and normally distributed, then Curve d.o.f. = 1 is the distribution of x_1^2. For this case the degrees of freedom or number of dimensions in which variability takes place = 1 and the distribution of x_1 is normal. Curve d.o.f. = 2 is the distribution of $(x_1^2 + x_2^2)/2$. Curve d.o.f. = 3 is the distribution of $(x_1^2 + x_2^2 + x_3^2)/3$. Etc. The curve for 30 or more degrees of freedom is not, on the scale used, visually distinguishable from a normal curve.

significance of means, differences of means, and of regression coefficients, for small samples,—say N less than 15. It is the distribution of these statistics computed from small samples drawn from a parent normal distribution. The χ^2 distribution, shown in Chart VIII III, is that of the squares of independent variables,—all under the assumption that these variables separately are normally distributed. The highly serviceable distribution of the ratio of two variances is another derivative of normally distributed measures.

Thus, in practical statistics, the normal distribution directly serves many situations and, by being the parent from which other distributions are derived, indirectly serves the field of contingency, analysis of variance, small sample theory, and still many other fields of advanced statistics.

SECTION 4. THE NORMAL DISTRIBUTION AS RELATED TO THE DESIGN OF EXPERIMENTS

In general, when an experiment is drawn up, simplicity and directness of treatment and interpretation will be served if

(*a*) the crucial variable can be analyzed into independent rather than correlated parts;

(*b*) where correlations exist, regressions, as explained in Chapter XI, are linear, not curvilinear;

(*c*) distributions are either rectangular or normal, not asymmetrical, bimodal, or of unusual kurtosis;

(*d*) residual errors are normally distributed, or, as in the case of t, χ^2, and F, the variance ratio, distributed as some function of a normal distribution.

Not infrequently a mathematical transformation of an initially non-normal variable into one normally distributed will, at the same time, make (a) possible and bring about (b). Thus in all four connections the normal distribution plays an important and frequently crucial role.

SECTION 5. DERIVATION AND TABLES OF THE NORMAL CURVE

There are innumerable ways of deriving the equation of the normal curve. Since all of these reveal it to be the limit of some process, the mathematical concepts of infinite series, limits, and of the calculus seem necessary to a complete understanding of its derivation. It has frequently been shown that the successive terms of the binomial $(p+q)^n$ yield the distribution of events when random samplings are drawn and also when errors of observation or of measurement are of a chance nature. It has also frequently been shown that this distribution, when N becomes large, approaches a normal distribution.

The accompanying derivation, which is that of the simplest of the Pearson system of curves, nicely illustrates the mathematical simplicity of the normal curve.

If, at any point in a continuous curve $y=f(x)$ a tangent to the curve is drawn, the slope, or tangent of the angle which it makes with the x axis, is approximately $\delta y/\delta x$,—the ratio of the small change taking place in y attendant upon a small change in x. The slope is exactly this ratio as the change in x, viz., δx, is made infinitely small. The calculus notation is

$$\text{slope} = \frac{dy}{dx}$$

We will observe that the slope properties of a symmetrical, unimodal curve, origin at the mean, whose tails approach ever more closely to the x axis, are that the slope equals zero when $x = 0$

and also when $y = 0$ (which y does when $x = \infty$, or $-\infty$). Algebraically, the simplest slope equation, called a differential equation, having these properties is

$$\frac{dy}{dx} = -cxy \qquad [8:01]$$

in which c is some positive constant. Corresponding to this equation expressing the slope properties is another equation, its integral, giving the relationship between y and x. This equation is

$$y = k\,e^{-c\frac{x^2}{2}}$$

in which k is some constant and e is 2.71828..., the Naperian base of logarithms. If we now choose c so that the standard deviation of the distribution is equal to 1.00, and so choose k that the total area under the curve is equal to 1.00, and substitute z for y merely to avoid certain ambiguities in notation which would otherwise arise, we have

$$z = \frac{1}{\sqrt{2\pi}}\,e^{\frac{-x^2}{2}} \qquad \text{Equation of unit normal distribution} \qquad [8:02]$$

The area to the left of any point x is designated p and is given by the integral of [8:02]. Thus

$$p = \int_{-\infty}^{x} z\,dx = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\,e^{\frac{-x^2}{2}}\,dx \qquad \text{Normal probability of a measure being less than } x \qquad [8:03]$$

Of course q $(= 1 - p)$ is the area to the right of x and is the probability of a measure being greater than x. Since the curve is symmetrical

it is only necessary to table x and z values for p between .5 and 1.0, or q between .5 and 0.0, as has been done in Table XV C. For x and z corresponding to a p value less than .5 one enters the table with the complement of p, viz., with q, and finds the value of z and, after attaching a minus sign, of x.

The two most common requirements of practical statistics are for values of x and z, knowing p, and for values of z and p, knowing x. Kelley's extensive *Table of Normal Probability Functions* (1947) gives directly x and z to eight decimal places for an argument of p of four decimal places. Entering Kelley's table with x yields p and z, but not as conveniently as, e.g., does Sheppard's Table (Pearson, 1914) in which the argument is x, to two decimal places, and the tabled entries are p [called in his notation $.5(1 + \alpha)$] and z to seven decimal places.

Of the many other tables of the normal curve the early and truly colossal table, considering the mechanical aids available at the time, of James Burgess (1897-98) and the later extensive tables of the National Bureau of Standards (1941 and 1942) are especially notable.

If we modify [8:02] so that the standard deviation is σ instead of 1.0 and the area N instead of 1.0 and call the ordinate y, the equation is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \qquad \text{The normal equation with origin at the mean, number of cases } = N\text{, and standard deviation of } \sigma \qquad [8:04]$$

If X is a raw score and M the mean, then $x = X - M$ and we may substitute $X-M$ for x in [8:04] and obtain the equation expressed in terms of raw scores.

SECTION 6. CERTAIN PROPERTIES OF THE NORMAL DISTRIBUTION

The exponential nature of the normal equation would seem, to the uninitiated, to suggest that its manipulation would be somewhat involved, but this is not the case because the intrinsic relationships maintaining in the curve are extremely simple, being equaled in simplicity only by those inherent in the rectangular and in the Poisson distributions, and because very adequate tables of the normal probability functions are available to the statistician. Just as logarithms became a daily device when adequate tables were published, so the normal curve has become such a tool because of the tabling of its functions.

We let μ_n' be the n'th moment from the raw score origin and μ_n the n'th moment from the mean. The successive moments of the normal distribution [8:04] are as follows:

$$\mu_0 = \int_{-\infty}^{\infty} y\, dx = N, \text{ the number of cases} \qquad [8:05]$$

Sometimes μ_0 is defined as $\frac{1}{N}\int_{-\infty}^{\infty} y\, dx$ in which case it of course = 1.0.

$\mu_1' = M$, the mean.

$$\mu_1 = \frac{1}{N}\int_{-\infty}^{\infty} y\, x\, dx = 0; \text{ and every other odd moment } = 0. \qquad [8:06]$$

The mean deviation $= \frac{2}{N}\int_{0}^{\infty} y\, x\, dx = \sqrt{\frac{2}{\pi}}\,\sigma$

(see . . . . [8:20]). . . [8:07]

$$\mu_2 = \frac{1}{N}\int_{-\infty}^{\infty} y\, x^2\, dx = V = \sigma^2 \qquad [8:08]$$

$$\text{Mean } |x^3| = \frac{2}{N} \int_0^\infty y\ x^3\ dx = \sqrt{\frac{8}{\pi}}\,\sigma^3 \ldots\ldots\ldots \quad [8{:}09]$$

$$\mu_4 = \frac{1}{N} \int_{-\infty}^{\infty} y\ x^4\ dx = 3V^2 \ldots\ldots\ldots \quad [8{:}10]$$

$$\text{Mean } |x^5| = \frac{2}{N} \sqrt{\frac{128}{\pi}}\,\sigma^5 \ldots\ldots\ldots \quad [8{:}11]$$

Every even moment is derivable from V and the even moment, which is two less than the original moment, through the relationship

$\mu_n = (n-1)\ V\ \mu_{n-2}$, wherein n is even [8:12]

so that

$$\mu_6 = 15\ V^3 \ldots\ldots\ldots \quad [8{:}13]$$

$$\mu_8 = 105\ V^4 \ldots\ldots\ldots \quad [8{:}14]$$

$$\mu_{10} = 735\ V^5,\ \text{etc.} \ldots\ldots\ldots \quad [8{:}15]$$

$\beta_1 = \dfrac{\mu_3^2}{V^3} = 0,$ Normal skewness (Pearson) [8:16]

$\beta_2 = \dfrac{\mu_4}{V^2} = 3,$ Normal kurtosis (Pearson) [8:17]

$\gamma_1 = \tilde{\mu}_3 / \sqrt{\tilde{\mu}_2^3} = 0,$ Normal Skewness (Fisher) [8:18]

γ_1 is a theoretical value and Fisher's sample estimate of it is called g_1.

$\gamma_2 = \dfrac{\tilde{\mu}_4}{\tilde{V}^2} - 3 = 0,$ A measure of normal kurtosis (Fisher) [8:19]

γ_2 is a theoretical value and Fisher's sample estimate of it is called g_2.

The equations just given for μ_0, μ_1, μ_1, and μ_2, hold for all distributions, those for the odd moments for all symmetrical distributions, and that for μ_4 for all mesokurtic distributions. The normal equation is that particular symmetric mesokurtic distribution for which the reduction equation [8:12] holds.

An approximate plotting of a normal curve can readily be made from the values at five points, as follows:

x	-2.50	-1.00	.00	1.00	2.50
z	.0175	.24	.40	.24	.0175
Ratio of z to z at the mean	$\frac{1}{25}$	$\frac{3}{5}$	1.00	$\frac{3}{5}$	$\frac{1}{25}$

When plotting, the slope is, of course, to be made zero at $x = 0$, and the points of inflexion of the curve are at -1 (i.e., $-1\ \sigma$) and 1 (i.e., $+1\ \sigma$).

The relationships between the standard deviation and other measures of variability are, in the normal distribution, as given herewith:

$$\text{A.D.} = \sqrt{\frac{2}{\pi}}\,\sigma = .79788456\ \sigma \quad . \ . \ . \ . \ . \ . \ . \quad [8:20]$$

$$Q = .67448975\ \sigma \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad [8:21]$$

$$Pv = P_{.93} - P_{.07} = 2.95158206\ \sigma \quad . \ . \quad [8:22]$$

The very slightly more precise normal optimal interpercentile range is

$$P_{.9308395} - P_{.0691605} = 2.9641450\ \sigma \quad . \ . \ . \ . \quad [8:22a]$$

The relative merit of these measures of vari-

ability, for samples drawn from normal populations, is shown in Table VI G.

The most common use to which the normal distribution is put is in finding the probability of a measure exceeding a certain value. This is given by the proportion of cases of greater value. It is necessary to express the certain value in which interested as a standard score. If the value is x, we compute the standard score

$$x = \frac{X-M}{\sigma}$$ A standard score (also called a z score) . . . .[8:23]

Table VIII III gives q, the proportion of cases greater than x for a few selected values of x.

TABLE VIII C

NORMAL PROBABILITIES

x STANDARD SCORES	q,-PROBABILITY OF EXCEEDING x	$2q$,-PROBABILITY OF EXCEEDING x OR FALLING SHORT OF $-x$
.0	.50	1.0
.67449	.25	.50
1.	.1587	.3173
1.64485	.05	.10
1.95996	.025	.05
2.	.02275	.0455
2.32635	.01	.02
2.57583	.005	.001
3.	.00135	.00270
3.09023	.001	.002
3.71902	.0001	.0002
3.89059	.00005	.0001

Since errors of observation are, in general, distributed in a nearly normal manner, an error in excess, in absolute amount, of $.67449\sigma$ is as likely to occur as an error less, in absolute amount, than $.67449\sigma$. For this reason, since its first use by Bessel and Gauss in 1815 and 1816,

the magnitude $.67449\sigma$ has been called a "probable error."

$$\text{P.E.} = .6744898\ \sigma \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad [8:24]$$

The precision of a statistic may be indicated by attaching the probable error to it thus: 14.7 ± 1.2. An alternative procedure is to give its standard deviation, called a standard error when the deviation in question is conceived of as an error, thus: 14.7 (st.er. = 1.8). Since this second procedure is arithmetically simpler, as it does not involve multiplication by .67449, and is more accurate, as it does not by implication assume a normal distribution of errors, it will be used in this text. The designation of the standard error by the ± sign, though found, is definitely bad practice, in that this sign has earlier and systematically been employed in connection with the probable error.

The probable error and the quartile deviation have been confused. $Q = (Q_3 - Q_1)/2$, one-half the difference between the first and the third quartiles in any distribution, and P.E. $= .67449\sigma$ in any distribution. Q and P.E. happen to be equal in the case of the normal distribution, but in general they are not equal and should not be used interchangeably.

We may note three degrees of refinement in judging of the precision of a statistic: *(a)* reporting its P.E. and thus by implication assuming a normal distribution of errors; *(b)* reporting its standard error, thus not stipulating the form of distribution of errors, and *(c)* reporting the presumptive form of distribution of errors (perhaps a *t*-distribution, a chi-square distribution, or a variance ratio distribution). The third procedure is highly meritorious, and especially so when small samples (or small number of degrees of freedom) are involved, but most problems of significance of a statistic are

answered by knowledge of its standard error, so the student should become thoroughly familiar with the computation of the standard deviations of all sorts of statistics.

The ratio of a statistic to its standard error is called a critical ratio.

In medical and chemical situations it is conceivable that some virus, bacterium, or other substance, may have an affect that is not proportional to the frequency of the substance, but to some power of the frequency. This suggests that we study the distribution of y^r, where r is some positive integral or fractional magnitude and y is the ordinate in a normal frequency curve.

$$y = \frac{N}{\sigma\sqrt{2\pi}} \exp - \frac{x^2}{2\sigma^2}$$

$$y^2 = \frac{\dfrac{N^2}{2\sigma\sqrt{\pi}}}{(\sigma\sqrt{.5})\sqrt{2\pi}} \exp \frac{-x^2}{2(\sigma\sqrt{.5})^2}$$

which is a normal distribution with variance $.5\,\sigma^2$ and number of cases $N^2/(2\,\sigma\sqrt{\pi})$.

$$y^r = \left(\frac{N}{\sigma\sqrt{2\pi}}\right)^r \exp \frac{-r\,x^2}{2\sigma^2} \qquad [8:24a]$$

This also is a normal distribution.

SECTION 7. STATISTICAL CONSTANTS DESCRIPTIVE OF PORTIONS OF A NORMAL DISTRIBUTION

CHART VIII IV

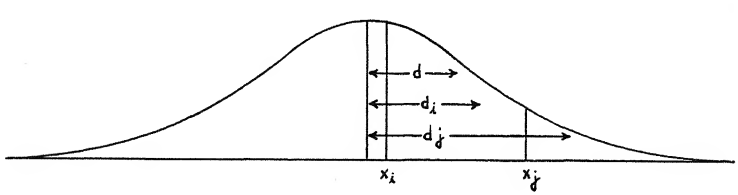


Chart VIII IV represents a unit normal distribution. The magnitude x is a standard score, the height of the curve at this point x is z, and the proportion of the total area above this point is q. The mean of the measures above x_j is d_j, and is approximately the distance labeled d_j. The mean of the tail above x_i is d_i and the mean of the portion between x_i and x_j is d_{ij}. Knowing q, the magnitudes x and z are gotten from a table of normal probability functions. A brief table of these functions is given in Table XV C.

A formula giving d, the mean deviation of a tail portion, is readily derived by means of the calculus

$$d = \frac{1}{q}\int_x^\infty xz\ dx = \frac{z}{q}$$ The mean deviation of a tail portion of a unit normal distribution [8:25]

If the tail is greater than one-half the curve, the formula becomes

$$d = \frac{z}{p} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \qquad [8:26]$$

The sign of d is positive if the tail is toward the right and negative if toward the left, when the x scale is from left to right.

Having a d, a distance in a unit normal distribution, the corresponding deviation from the mean if the standard deviation is σ is simply $d\sigma$.

Let q_j be the proportion to the right of x_j, q_i the proportion to the right of x_i, q_{ij} the proportion in the interval from x_i to x_j, i.e., $q_{ij} = (q_i - q_j)$. Since the mean of a sum is the average of the means of the parts, each weighted according to the number of cases, we have

$d_i q_i = d_{ij}(q_i - q_j) + d_j q_j$, which with [8:25]

yields

$$d_{ij} = \frac{z_i - z_j}{q_i - q_j} = \frac{z_i - z_j}{q_{ij}} \quad \text{The mean deviation of a portion of a unit normal distribution} \qquad [8:27]$$

As employed in this text q is generally a proportion less than .5, but in [8:27] it is the proportion to the right of the point of dichotomy and may be less or greater than .5. When so employed the sign of d_{ij} is correctly given by the sign of $(z_i - z_j)$.

Table VIII D herewith illustrates the transformation of an ordered variable,—teachers' marks A, B, C, D, E, F,—into quantitative values, under the assumption that the talent represented by the ordered data is normally distributed. The mark A corresponds to a position 1.692 standard deviations above the mean, etc.

TABLE VIII D

NORMALIZING AN ORDERED VARIABLE

MARKS	PERCENTAGE OF PUPILS RECEIVING MARK INDICATED	FROM TABLE XV C				$d_{ij} = \frac{z_i - z_j}{q_{ij}}$
		q_i	q_j	z_i	z_j	
A	11.4	.114	.000	.192900	.000000	1.692
B	34.7	.461	.114	.397034	.192900	.588
C	32.5	.786	.461	.291399	.397034	−.325
D	10.2	.888	.786	.190478	.291399	−.989
E	9.0	.978	.888	.052485	.190478	−1.533
F	2.2	1.000	.978	.000000	.052485	−2.386

It is frequently desirable to have further statistics of portions of a normal distribution. We herewith give the calculus statement for obtaining the variance, the third moment and the fourth moment of a portion bounded by x_i and x_j. The calculus formulas are given to illustrate the derivation, but are not needed in the practical

utilization of the method.

The mean of the squared deviations of the q_{ij} measures in the interval x_i to x_j, from the mean of the entire distribution is V'_{ij}, and from their own mean, d_{ij}, it is as given by [8:29]

$$V_{ij} = V'_{ij} - d^2_{ij}$$

$$V'_{ij} = \frac{1}{q_{ij}} \int_i^j x^2 \, z \, dx \qquad \text{[8:28]}$$

$$= 1 + \frac{x_i z_i - x_j z_j}{q_{ij}}$$

*Mean squared deviation, from total distribution mean, of a portion of a unit normal distribution

Accordingly

$$V_{ij} = 1 + \frac{x_i z_i - x_j z_j}{q_{ij}} - d^2_{ij} \qquad \text{[8:29]}$$

Variance of a portion of a unit normal distribution

By very similar procedures we obtain the third and fourth moments of portions:

$$\mu'_{3:ij} = \frac{1}{q_{ij}} \int_i^j x^3 \, z \, dx$$

$$= \frac{x_i^2 z_i - x_j^2 z_j}{q_{ij}} + 2\, d^2_{ij} \qquad \text{[8:30]}$$

Mean cubed deviation, from total distribution mean, of a portion of a unit normal distribution

* Having p, q, x, and z as defined in connection with a unit normal distribution, we have,

$$dp = z \, dx \text{ [8:28a], and}$$
$$dz = -xz \, dx \text{ [8:28b].}$$

The various integrals here needed are then obtained by integrating by parts and evaluating at the limits.

$$\mu_{3:ij} = \mu'_{3:ij} - 3V'_{ij}\, d_{ij} + 2d^3_{ij}$$

$$= \frac{x_i^2 z_i - x_j^2 z_j}{q_{ij}} + (2 - 3V_{ij})\, d_{ij} - d^3_{ij}$$

Third moment of a portion of a unit normal distribution [8:31]

$$\mu'_{4:ij} = \frac{1}{q_{ij}} \int_i^j x^4\, z\, dx$$

$$= \frac{x_i^3 z_i - x_j^3 z_j}{q_{ij}} + 3\, V'_{ij}$$

Mean fourth powered deviation, from total distribution mean, of a portion of a unit normal distribution [8:32]

$$\mu_{4:ij} = \mu'_{4:ij} - 4\, \mu'_{3:ij}\, d_{ij} + 6\, V'_{ij}\, d^2_{ij} - 3\, d^4_{ij}$$

$$= \frac{1}{q_{ij}} [x_i^3 z_i - x_j^3 z_j - 4(x_i^2 z_i - x_j^2 z_j)\, d_{ij}$$

$$+ 6(x_i z_i - x_j z_j) d^2_{ij} + 3(x_i z_i - x_j z_j)]$$

$$+ 3 + 6d^2_{ij} - 8d^3_{ij} - 3d^4_{ij}$$

Fourth moment of a portion of a unit normal distribution [8:33]

SECTION 8. THE SIGNIFICANCE OF DIFFERENCES BETWEEN THE MEANS OF TAIL PORTIONS OF A NORMAL DISTRIBUTION

Let us be given a sample of N cases upon each of which has been made a measurement, X, with, so far as we can tell, the same precision throughout, so that the standard error of measurement, s, may be taken to be constant for all measures. The variance, due to the error inherent in the technique of measurement, of the mean of any subsample of N_i cases is s^2/N_i and is not a function of the distribution of these N_i cases, nor of the distribution of the N cases.

Let the mean of an upper tail portion of N_i cases be D_i and the mean of a lower non-overlapping tail portion of N_j cases be D_j. The difference between the means is (D_i-D_j) and the variance error of this difference is $(\frac{s^2}{N_i}+\frac{s^2}{N_j})$. The critical ratio of the difference divided by its standard error is $(D_i-D_j)/(s\sqrt{\frac{1}{N_i}+\frac{1}{N_j}})$, where the error in question is that due to measurement and not due to sampling.

If the sample is normally distributed, we readily can find the size of the tail portions which make this critical ratio a maximum. It is axiomatic that the upper and lower portions will be equal. Let the X's be expressed in terms of standard scores. Let x_i be the point of dichotomy for the upper tail, q_iN the number of cases in the tail, and d_i as given by [8:26] the mean of the tail. For the lower tail the point of dichotomy is $-x_i$, the number of cases q_iN, and the mean $-d_i$. The critical ratio of difference between means divided by standard error is

$$\frac{2\,d_1}{s\sqrt{\dfrac{2}{q_1 N}}} = \frac{\sqrt{2N}}{s}\, d_1\,\sqrt{q_1} = \frac{\sqrt{2N}}{s}\left(\frac{z_1}{\sqrt{q_1}}\right)$$

in which N and s are constants. An examination of a table of the normal probability functions z and q enables us to find the value of q for which this function is a maximum. It is the value for which $q = z/(2x)$, i.e., $q = .2702678$.

Thus when upper and lower groups are selected in order to bring out most decisively the difference between means of the upper and lower, the two groups should consist of 27 per cent each, if the distribution for the sample entire is normal.*

SECTION 9. FITTING A NORMAL CURVE TO DATA

The equation of the fitted curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}}\, e^{\frac{-(X-M)^2}{2V}} = \frac{N}{\sigma}\, z \quad . \; . \; . \; . \; . \quad [8.34]$$

in which z is the ordinate in a table of the unit normal distribution corresponding to the deviate $x[= (X-M)/\sigma]$. It is frequently best to plot this curve for values of X which are class indexes, because then a precise comparison can be made between the computed curve and the actual data. The value of y given by [8:34] is the

* A fuller proof and further aspects of this matter have been discussed by Truman L. Kelley. "The selection of upper and lower groups for the validation of test items," J. ED. PSYCH., Vol. 30, no. 1, Jan. 1939.

frequency per unit interval. If the data have been grouped, using an interval i, the frequency called for is

$$iy = (i\ N\ z)/\sigma \quad . \quad . \quad . \quad . \quad . \quad . \quad . [8:35]$$

We will illustrate the steps in the computation, using the New York daily maximum temperature data of Chapter IV, Table IV B. The calculation of the mean and standard deviation, using a grouping interval of 3°, has already been made in Chapter VI, Section 6. The M = 81.7742 and σ = 6.2358. In Table VIII E herewith we give the class indexes, the equivalent standard scores, x_j, for the class indexes, the corresponding or equivalent ordinates, z_j, in the unit normal distribution, these multiplied by Ni/σ providing the theoretical frequencies, $\tilde{f}_j$, in the classes, which are to be compared with the actual frequencies, f_j, as shown. Though a grouping of a number of small theoretical frequency classes is called for, the actual grouping here employed for the upper and lower tail portions is slightly coarser than is desirable, as explained in Chapter IX, Section 5.

$$x_j = \frac{X_j - M}{\sigma} = \frac{X_j - 81.7742}{6.2358} = .16036\ X_j \quad - 13.1137$$

$$\tilde{f}_j = \frac{z_j\ N\ i}{\sigma} = \frac{z_j\ 62 \times 3}{6.2358} = 29.8278\ z_j$$

The difference between f_j and $\tilde{f}_j$ is the cell divergence; $(f_j - \tilde{f}_j)^2/\tilde{f}_j$ is the cell square contingency; and the sum of these for all cells is the square contingency, χ^2. Chi-square is the sta-

tistic which provides a test of the goodness-of-fit of the observed data to the theoretical curve.

TABLE VIII E

COMPUTATION AND FREQUENCIES FOR NORMAL CURVE FITTED TO MAXIMUM TEMPERATURE DATA OF TABLE IV G

CLASS INDEXES X_j	x_j	z_j	$\tilde{f}_j$	$\tilde{f}_j$	f_j	$\frac{(f_j - \tilde{f}_j)^2}{\tilde{f}_j}$
60	-3.49181	.000898	.03			
63	-3.01071	.00429	.13			
66	-2.52962	.01627	.49	5.59	4	.45
69	-2.04853	.04894	1.46			
72	-1.56743	.11679	3.48			
75	-1.08634	.22113		6.60	6	.05
78	- .60525	.33217		9.91	5	2.43
81	- .12415	.39587		11.81	23	10.60
84	.35694	.37432		11.17	13	.30
87	.83803	.28081		8.38	6	.68
90	1.31913	.16713		4.99	1	3.19
93	1.80022	.07892	2.35			
96	2.28131	.02957	.88			
99	2.76240	.00879	.26	3.56	4	.05
102	3.24350	.00207	.06			
105	3.72455	.000388	.01			
				62.01	62	17.75

$\chi^2 = 17.75$; degrees of freedom $= 5$; $\dfrac{\chi^2}{\text{d.o.f.}} = 3.55$

$$F_{5\,\infty} = \frac{3.55}{1}; \text{ and } \Leftrightarrow P = .0034$$

The frequencies of the observed distribution for the successive classes are recorded in the f_j column, and the theoretical frequencies in the f_i column.

This is not the chapter in which fully to discuss χ^2, degrees of freedom, and the goodness-

of-fit test, but it will be noted that since the test involved eight class frequencies and since the theory accepted three constants determined from the data, namely, N, the mean, and the standard deviation,* there are eight minus three, or five, degrees of freedom. Thus $(\chi^2/5)/1$ is $F_{5\infty}$ a variance ratio with five degrees of freedom in the numerator variance and an infinite number in the denominator. By methods given in Chapter IX, an equivalent P, the probability that a divergence as great as that observed may be attributed to chance, is found and this constitutes the final evidence as to goodness-of-fit.

It would commonly be stated that the theory being tested is whether the deviations from normality present in the observed data are of such magnitude as to be attributed to chance, if the observed 62 items were randomly drawn items from a normal parent population. This is not quite accurate, for actually the theory being tested concerns itself with that infinite number of subsamples of 62 each of a parent population for which the mean is 81.7742 and the standard deviation is 6.2358. These sub-samples do themselves constitute an infinite population and the theory being tested is whether the observed sample is a chance deviate from such sub-samples. For present purposes the distinction is important because it accounts for the loss of three degrees of freedom.

For the data in hand, with P = .0034, we see that there is less than 1 chance in 100 that the theory of normality is tenable. Knowing nothing about the source of the data we would draw this conclusion, but knowing the source and knowing that successive daily maximum temperatures at a single spot are correlated and not independent,

* $V = (\Sigma fx^2)/N$ is a linear function of the f's.

we have a ready explanation in that our items are not independent random samplings. Prior to the study we would, in fact, have cogent grounds for expecting the distribution to be non-normal, even though we do not know just what form to expect.

SECTION 10. THE DISTRIBUTION OF THE SUM OF NORMAL VARIABLES

Let x_1, x_2 each be deviations from means, normally distributed, and independent of each other, with variances V_1 and V_2, and let $x=x_1+x_2$. We shall determine the moments of the distribution of x.

$$x^k = x_1^k + k\, x_1^{k-1}x_2 + \frac{k(k-1)}{1\cdot 2} x_1^{k-2}x_2^2 + \cdot\ \cdot\ \cdot$$

In getting the k'th moment of x we observe that all summations involving x_1 or x_2 to an odd power vanish because the variables are independent and the summation of odd powers of each separately is zero; and summations involving even powers, g and h, can be expressed in parts, thus

$$\Sigma x_1^g x_2^h = N\, \mu_{g,1}\mu_{h,2}$$

Finally, utilizing [8:12] in connection with the moment of x_1 and x_2, we obtain after making the necessary summations and substitutions, in case k is even, the following k'th moment of x:

$$\mu_k = (k-1)\ (k-3).\ .\ .\ (1)(V_1 + V_2)^{k/2}$$

Of course

$$\mu_{k-2} = (k-3).\ .\ .\ (1)(V_1 + V_2)^{(k-2)/2}$$

Since $\mu_2 = V_1 + V_2$ we obtain

$$\mu_k = (k-1)\mu_2\mu_{k-2}$$

This is [8:12], the fundamental relationship between the even moments in a normal distribution. We have thus proven that x, the sum of two normally distributed independent variables, is normally distributed. It is noteworthy that this maintains even though the variables have unequal variances.

If x is the sum of three normally distributed independent variables, then x is normally distributed. The proof is immediate, for we can first combine x_1 and x_2, getting a normally distributed variable, and this combined with x_3 yields a normally distributed variable.

Finally, *the sum of any number of independent normally distributed variables is normally distributed.*

If we let $x_{i'} = -x_i$ and substitute $-x_{i'}$ for x_i in the sum, the preceding statement obviously holds, The proof that adding a constant, or multiplying a variable by a constant, does not alter the relationship is equally simple, so we can assert: *Any linear combination of normally distributed independent variables is normally distributed.*

SECTION 11. THE RELATIONSHIP OF CHI-SQUARE, T-SQUARE, AND F TO THE NORMAL DISTRIBUTION

The purpose of this section is to state and illustrate, but not to prove these relationships.

Let Δ be a magnitude of the small order of the differentials of calculus. Let

$$p_{\Delta_1} = \frac{1}{\sigma_1\sqrt{2\pi}} \exp \frac{-x_1^2}{2V_1}\,\Delta_1$$

This equals the probability of a measure falling in the interval x_1 to $x_1 + \Delta_1$. Similarly for a second variable

$$p_{\Delta_2} = \frac{1}{\sigma_2\sqrt{2\pi}} \exp \frac{-x_2^2}{2V_2} \Delta_2$$

If x_1 and x_2 are independent, the probability of the joint occurrence, or of obtaining a case falling within the Δ_1 interval for the first variable and within the Δ_2 interval for the second variable is

$$p_{\Delta_1} p_{\Delta_2} = \left[\frac{1}{\sigma_1\sigma_2\ (2\pi)} \exp \frac{-1}{2}\left(\frac{x_1^2}{V_1} + \frac{x_2^2}{V_2}\right)\right] \Delta_1\Delta_2$$

The distribution in two-dimensional space of the function within the brackets is related to a χ^2 distribution with two degrees of freedom, and in general the i dimensional distribution of

$$\frac{1}{\sigma_1\sigma_2 \ \ldots \ \sigma_1 (2\pi)^{\frac{i}{2}}} \exp \frac{-1}{2}\left(\frac{x_1^2}{V_1} + \frac{x_2^2}{V_2} + \ \ldots \ + \frac{x_i^2}{V_i'}\right)$$

is related to a χ^2 distribution with i degrees of freedom. The use of V_i' instead of V_i is to avoid ambiguity with the more usual meaning of V_i as given in [8:38].

On the two-dimensional surface having dimensions x_1/σ_1 and x_2/σ_2, the distance from the origin of any point is $\sqrt{(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2}$, a one-dimensional variable, which we designate χ, chi. Of course it > 0.

$$\chi^2 = \frac{x_1^2}{v_1} + \frac{x_2^2}{v_2}$$

Since x_1 and x_2 are uncorrelated the mean of this χ^2, if many samples are taken is

$$\text{Mean } \chi^2 = V(\frac{x_1}{\sigma_1}) + V(\frac{x_2}{\sigma_2})$$

$$= 1 + 1 = 2 = \text{the number of d.o.f.}$$

For the general case we have

$$\text{Mean } \chi^2 = i \quad \text{Mean of } \chi^2 \text{ with } i \text{ d.o.f.} \quad . . . . [8:36]$$

Since the mean $\chi_i^2/i = 1$ (the median < because of skewness of the distribution), any set of x_1, x_2, . . . x_3 values yielding a $\chi_i^2/i > 1$ (more precisely greater than the median) tends to indicate variability factors in χ_i^2 in excess of those postulated. We have

$$\chi_i^2 = \frac{x_1^2}{V_2} + \frac{x_2^2}{V_2} + . . . + \frac{x_i^2}{V_i'} \quad \text{Chi-square with } i \text{ degrees of freedom} \quad . . . [8:37]$$

From the structure shown in [8:37], χ_i^2/i is clearly a variance, as is also χ_i^2 itself. To simplify future notation we write

$$V_i = \chi_i^2 \quad . . . . [8:38]$$

This V_i is the quantity entering into a variance ratio

$$F_{ij} = \frac{V_i/i}{V_j/j} \quad \text{A variance ratio based upon } i \text{ (numerator) and } j \text{ (denominator) d.o.f.} \quad . . . . . . \text{See}[9:21]$$

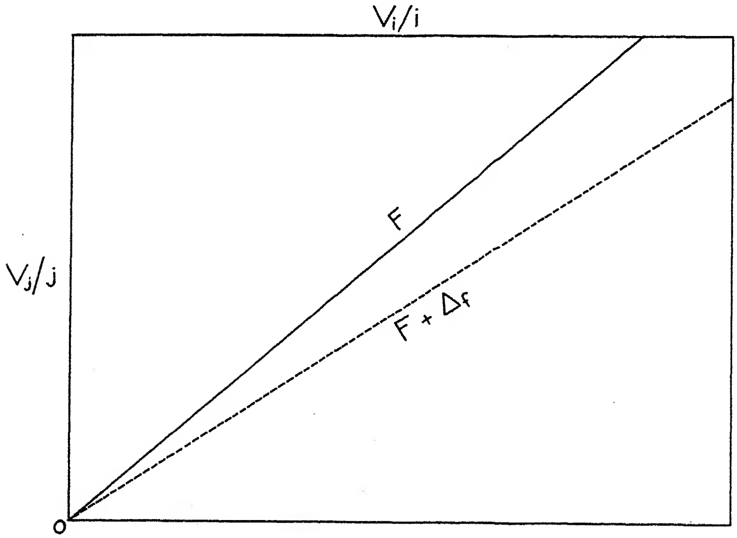
As a limiting case a χ_i^2/i is a variance ratio in which the d.o.f. of the denominator variance is ∞, for $V_\infty/\infty = 1$.

Accordingly

$$F_{1\infty} = \frac{V_i}{i} = \frac{\chi_i^2}{i} \qquad \text{Chi-square expressed as a variance ratio} \qquad [8:39]$$

The variance ratio distribution is that of a quotient. The distribution of the numerator alone of F_{ij} is that of V_i/i, and of the denominator is that of V_j/j. Consider a scatter diagram Chart VIII V, with dimensions V_i/i and V_j/j. A fixed value of F is represented by a straight line passing through the point (0,0). $F + \Delta_f$ is another straight line slightly to the right of the former and the frequency of cases between the two lines is Δ_p. the probability of a variance ratio between F and $F + \Delta_f$. The entire frequency to the right of the line F is P, the probability of a variance ratio greater than F. Thus, though several steps removed, F is a derivative of the normal distribution.

CHART VIII V



The t_j^2 (student's "t" squared) distribution is an f_{1j} distribution, the distribution of a variance ratio having one degree of freedom in the numerator and j degrees in the denominator.

The x^2 (x of a unit normal distribution squared) distribution is the distribution of an $F_{1\infty}$, a variance ratio having one degree of freedom in the numerator and ∞ degrees of freedom in the denominator.

The F distribution spans the field. However, since the determination of P from F would involve a three-way table, the dimensions being F, i, and j, its complete tabulation would be a collossal undertaking. This, in a modified form has been done in the 500-odd pages of Pearson's *Tables of the Incomplete Beta-Functions* (1934). Snedecor (1938), Fisher and Yates (1938), and others have tabled F for selected values of P and extended values of i and j.

In Chapter IX we give a transformation of F voiding the need of any of these tables, but the use instead of a normal probability table to obtain P for any F.

CHAPTER IX

THE STATISTICS OF ATTRIBUTES

SECTION 1. SITUATIONS WHEREIN QUALITATIVE SERIES ARISE

Data are placed in categories when a quantitative relationship between the classes does not exist (more accurately "is not known to exist"), when the quantitative relationship is only vaguely surmisable, or when the known quantitative relationship between classes is neglected because the more primitive and simple qualitative methods would seem to suffice. Customary types of qualitative series would be such as a number of men classified by nationality, or by eye color, or by vocation, or by religion, or by marital status, or by language spoken, etc. Handling such situations involves the statistics of attributes, or of qualitative series, or of categories. *The fundamental item is the frequency in a class.*

SECTION 2. THE FREQUENCY IN A CLASS

Let the classes into which the data can fall be a, b, c, . . . k and the observed frequencies be f_a, f_b, . . . f_k, and let f_s stand for any one of these frequencies as s takes values from

a to *k* inclusive. For many purposes it is desirable to distinguish between those situations in which the probability of a case falling in a class is known a priori and those in which it can only be surmised within limits from the observed proportions in the classes. If people are classified into six vocational groups and the observed frequencies, when 100 are drawn at random, are 18, 16, 22, 13, 14, 17, we can estimate the probability of a measure falling in a given class. We would estimate the chance of a person being in the first class as .18 and can use this estimate, with precautions to be noted, in judging whether a subsequent sample is one drawn from the same population. This is the most usual qualitative series.

Another, not differing in face appearance, but actually differing exists when one knows, because of antecedent knowledge, the probabilities of a measure falling in the different classes. Such a series is a stochastic series. If a homogeneous cube with red, yellow, blue, violet, orange, and green faces, is tossed 100 times the frequency of occurrence of the various faces up might be

Color	red	yellow	blue	violet	orange	green
f	8	16	22	13	14	17

We can make tests of this series which are not possible with the other for the a priori probabilities are 1/6 for each class.

Had the cube been a die with 1, 2, 3, 4, 5 and 6 pips upon its faces, the frequencies of occurrence would have the same fundamental properties as before, though the X score is now a quantitative one. For some purposes the number of pips should be treated merely as categories, but for other purposes, as for example when paying gambling debts, they must be treated as quantitative measures. In the series involving the

cube with colored faces some genius might be able to see a quantitative scale and use the results quantitatively, but the ordinary non-physicist could only treat it as a categorical series. This observation is made because it is true that series which seem to be merely qualitative may frequently be found, with further study, to be quantitative in a respect that has scientific and social consequences. A good example of this in psychology is in the quantitative measures that have been teased out of qualitative expressions of interests and attitudes. The enrichment of understanding is so great when qualitative data can be recast into a quantitative mold that the idea of doing so should be the initial thought of a student when first confronted with a qualitative series.

We will first derive a few statistics pertaining to a stochastic variable. Let us know a priori that the probability of a measure drawn at random in class *A* is p and that the probability of its not so falling is q, —$(p+q=1)$. Let X = the number found in class *A* when a sample of N is drawn. If $N = 1$ then $X = 0$ or 1. The theoretical, or true, distribution would be the average of many such samples, and would have mean and moments as in Table IX A herewith.

TABLE IX A

COMPUTATION OF MOMENTS FOR A TWO-POINT DISTRIBUTION

X	PROBABILITY	f = PROB. NO. WHEN T SAMPLES OF 1 EACH ARE DRAWN	fx	fx^2	fx^3	fx^4
0	q	Tq	0	0	0	0
1	p	Tp	Tp	Tp	Tp	Tp
SUMS	. 1	T	Tp	Tp	Tp	Tp
MEANS	. . .	. . . .	p	p	p	p

M (Mean of X's) $= p$

$V = p - p^2 = pq$, by [6:05]

$\mu_3 = p - 3p^2 + 2p^3 = pq\,(q-p)$, by [6:47]

$\mu_4 = p - 4p^2 + 6p^3 - 3p^4 = pq(1-3p+3p^2)$, by [6:48]

In a similar manner we can compute the moments of the X distribution if T (a large number) samples of 2 each are drawn.

The chance that the first case $= 0$ is q
The chance that the second case $= 0$ is q
The chance that both cases $= 0$ is q^2

The chance that the first $= 0$
and the second $= 1$ is pq
The chance that the second $= 0$
and the first $= 1$ is qp
The chance that the one $= 0$
and the other $= 1$ is $2\,pq$

The chance that the first case $= 1$ is p
The chance that the second case $= 1$ is p
The chance that both cases $= 1$ is p^2

The sum of these chances $(p^2 + 2pq + q^2) = (p+q)^2 = 1$, as of course must be the case. When samples of N are drawn the successive probabilities are the successive terms of $(p+q)^N$. We could compute moments, as in Table IX A, having $X = 0$, 1 and 2, and probabilities q^2, $2pq$ and p^2, but will proceed immediately to the general case of the distribution of the frequencies in a class, for which the probability of a measure drawn at random being in the class is p, when T (a large number) samples of N each are drawn. The (fX_i) in column 3 is the entry under fX in column 2 for $X = i$, for example $(fX_1) = TNp\ [q^{N-1}]$.

X	f = PROB. NO. WHEN T SAMPLES OF N EACH ARE DRAWN	fX	fX^2
0	Tq^N	0	0
1	$TNq^{N-1}p$	$TNp[q^{N-1}]$	(fX_1)
2	$T\frac{N(N-1)}{1 \times 2}q^{N-2}p^2$	$TNp[(N-1)q^{N-2}p]$	$(fX_2)+TN(N-1)p^2[q^{N-2}]$
3	$T\frac{N(N-1)(N-2)}{1 \times 2 \times 3}q^{N-3}p^3$	$TNp[\frac{(N-1)(N-2)}{1 \times 2}q^{N-3}p^2]$	$(fX_3)+TN(N-1)p^2[(N-2)q^{N-3}p]$
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$
N	$T\frac{N \cdot \,.\,.(1)}{1 \times .\,.\, \times N}q^{N-N}p^N$	$TNp[\frac{(N-1).\,.\,.(1)}{1 \times \cdot\,\cdot \times (N-1)}p^{N-1}]$	$(fX_N)+TN(N-1)p^2[\frac{(N-2)\ldots(1)}{1 \times \ldots \times (N-2)}p^{N-2}]$
	T	$TNp(p+q)^{N-1} = TNp$	$TNp+TN(N-1)p^2(p+q)^{N-2}=TNp(1+Np-p)$
Means		Np	$Np(1+Np-p)$

TABLE IX B

COMPUTATION OF THE MOMENTS OF THE FREQUENCY IN A CLASS

Mean of the frequencies in the class $= Np$ [9:01]

Variance $= Np(1+Np-p)-(Np)^2$

$= Np(1-p) = Npq$ (see [14:21]) [9:02]

By similar calculation with fX^3 and fX^4, or more simply by utilizing [14:18], obtain

$\mu_3 = Npq(q-p)$ . . . . (see [14:22]) [9:03]

$\mu_4 = Npq[1+3(N-2)pq]$ (see [14:23]) [9:04]

$$\beta_1 = \frac{\mu_3^2}{V^3} = \frac{(q-p)^2}{Npq}$$ Skewness (Pearson) of frequency in a class [9:05]

$$\beta_2 = \frac{\mu_4^2}{V^2} = 3 + \frac{1-6pq}{Npq}$$ Kurtosis (Pearson) of frequency in a class [9:06]

The higher moments can be gotten by Romanovsky's reduction formula [14:18]. Thus the complete distribution of frequencies in a class, due to random sampling, is available.

Let us examine this distribution for different values of N and p.

If p is very small and N large then the moments, V, and μ_3 approach the value M, and the higher moments are given by [14:34], [14:35], [14:36], etc. This is a Poisson distribution. The particular fact that $M = V$ underlies the computation of χ^2 based upon frequencies in classes.

If N is large and p not very small then $\beta_1 \sim 0$, which is normal skewness, and $\beta_2 \sim 3$, which is mesokurtosis, and all higher moments approximate the normal distribution values. The approach to the normal distribution is most rapid if $p = q$.

For all values of N we note that $\beta_2 = 3$ when $(1-6pq) = 0$, i.e., when $p = .5 + \frac{1}{6}\sqrt{3} = .211325$, or .788675. The distribution is platykurtic for values of p between these two values. Otherwise it is leptokurtic, but the deviation from mesokurtosis is slight if N is even of moderate size.

We can use the binomial distribution to get a precise test of the significance of a difference between an observed and a theoretical class frequency. With no other information to help in a solution let us be given the fact that Joe matches a coin with John and wins eight times in ten. What is the likelihood that, if the coins are unbiased and the throws fair, so one-sided an outcome would arise as a matter of chance? The assumption is that $p = q = .5$, so the distribution of probabilities is that given by successive terms of $(.5 + .5)^{10}$, which are as follows:

TABLE IX C

CERTAIN BINOMIAL PROBABILITIES

Joe wins = X	0	1	2	3	4	5	6	7	8	9	10	
Probability of occurrence	.00097 65625	.00976 56250	.04394 53125	.11718 75000	.20507 81250	.24609 37500	.20507 81250	.11718 75000	.04394 53125	.00976 56250	.00097 65625	1.00000 00000

Thus the chance probability that Joe would win eight times in ten is .0439453125. The probability that he would have eight or more successes is .0546875 (= .0009765625 + .0097656250 + .0439453125). The probability that an outcome as extreme as this would arise as a matter of

chance is double this, or .109375. This last is the figure that ordinarily concerns one. It is the P of goodness-of-fit tests and states the chance that a situation as extreme as the one observed would arise as a matter of chance.

The usual way of obtaining an approximate P is by means of χ^2, the square (or squared) contingency. We will apply it to the present problem and compare the P therefrom with the correct answer, .109375.

SECTION 3. CHI-SQUARE

Let f_s = the observed frequency in the cell, or class, s. . . . . . . [9:07]

Let $\tilde{f}_s$ = the frequency in this cell according to the theory being tested. [9:08]

$f_s - \tilde{f}_s$ = the cell divergence. . . . . [9:09]

$(f_s - \tilde{f}_s)^2$ = the cell square divergence.

$$\frac{(f_s - \tilde{f}_s)^2}{\tilde{f}_s} = \chi_s^2 = \text{the cell square contingency} \quad [9:10]$$

$$\chi^2 = S\,\chi_s^2 = S\left[\frac{(f_s - \tilde{f}_s)^2}{\tilde{f}_s}\right] = \text{the sum of such for all cells} = \text{the square contingency} \quad [9:11]$$

The number of degrees of freedom in χ^2, as fully explained in Section 4, is here designated d.o.f. F_{ij} is a variance ratio with i degrees of freedom in the numerator and j in the denominator.

$$\frac{\chi^2}{\text{d.o.f.}} = F_{\text{d.o.f.}\infty} = $$ a variance ratio with d.o.f. degrees of freedom in the numerator and infinity degrees of freedom in the denominator

P from F is the chance that a situation deviating from the theoretical one as much as does the observed one would arise as a matter of chance. For the coin matching problem, P from χ^2 will be only an approximate answer because the χ^2/d.o.f. and the F distributions derive from the assumption that the elementary variable is normally distributed and reference to [9:05] and [9:06] show this not to be strictly true for frequencies in a class. The cell square contingency is an approximation to a critical ratio squared for, if p is small, $\tilde{f} = Np \sim N\,pq = V$. Wilson, Hilferty, and Maher* (1931) observe for a more extreme example than ordinarily occurs that it makes no particular difference whether Np, or Npq, is employed in the denominator when calculating the cell square contingencies. We may therefore think of the cell square contingency as very similar to a squared critical ratio. If these separately are normally distributed the sum of a number of such is precisely χ^2, as tabled. For the coin matching problem we have:

Number of Successes	Number of Failures
8	2
5	5
3	− 3
1.8	1.8

* Wilson, Hilferty, and Maher note that with k cells and $(k-1)$ d.o.f. $\chi^2/(k-1)$ and χ'^2/k, in which $\chi'^2 = S\left[\frac{(f_s \tilde{f}_s)^2}{N\,p_s q_s}\right]$, are very similarly distributed.

The four entries in each cell are, in order,

f_s = the observed cell frequency

$\tilde{f}_s$ = the theoretical cell frequency

$(f_s - \tilde{f}_s)$ the cell divergence

$$\frac{(f_s - \tilde{f}_s)^2}{\tilde{f}_s} = \chi_s^2 = \text{the cell square contingency}$$

$$\chi^2 = S\chi_s^2 = 3.6$$

The number of degrees of freedom is 1 for, with 10 matchings and 8 successes, the number of failures, 2, is dictated. χ^2/d.o.f. = 3.6/1 and from tables (either a χ^2 with 1 degree of freedom; a $F_{1.\infty}$ with 1 degree of freedom in the numerator and ∞ in the denominator; a t^2 with ∞ degrees of freedom in the denominator; or a x^2 of a normal distribution) we find P = .05775. This result is not close to the correct answer, but Yates (1938) has pointed out the need for a correction in χ^2, *in case of one degree of freedom. The correction is accomplished by "reducing by .5 the values which are greater than expectation and increasing by .5 those which are less than expectation."*

The intent of Yates' correction is to adjust for the fact that observed frequencies must be integers, while the theoretical distribution (the χ^2 distribution) is continuous. If, in a continuous distribution, a value such as 7.51 must be assigned an integral value it is called 8.0, which is to say that in fact the value 8.0 covers all values from 7.5 to 8.5.

If we therefore consider our observed number

of successes to be 7.5, and of failures 2.5 and recompute, we obtain $\chi^2 = 2.5$ and $P = .11385$ which is in fair agreement with the precise value (.109375).

SECTION 4. χ^2 FROM THE GENERAL CONTINGENCY TABLE

Let us now consider the problem wherein a priori probabilities for the different cells are unknown. If 100 people, drawn at random in city α, 200 drawn at random in city β, and 300 drawn at random in city γ fall into vocational groups as given in Table IX D, we can ask and answer the question "Are the discrepancies in the relative frequencies for these three series such as might readily arise as a matter of sampling if all three drawings are in fact from the same parent population, or do they exceed this so that we must believe that they come from different parent populations, i.e., that communities α, β and γ are of different sorts?"

We must compare these observed cell frequencies with theoretical cell frequencies, but we have no a priori values. However, on the assumption that the three communities are the same, we can assign theoretical values which are in agreement with the marginal totals N_s and N_t. To illustrate for cell γA: Half the cases are γ cases and 1/6 of them are A cases, so clearly 1/2 times 1/6 of them, or 1/12 of them, [(1/2)(1/6) 600 = 50], must, in the population, be γA cases. This, accordingly, is $\hat{f}_{\gamma A}$, the theoretical γA cell frequency. In general, $\frac{N_s}{N} \times \frac{N_t}{N} \times N$ is the theoretical cell frequency for the cell at the intersection of row s and column t, which we designate cell st.

The 600 people whose status is recorded in

Table IX D constitute a sample. They are, with certain limitations, to be thought of as randomly chosen from an infinite population. If there were no limitations, a second sample might not have 600 cases, or 100 from City α, 100 in Vocation A, etc., but we are not interested in the facts that there were 600 cases, 100 from City α, 100 in Vocation A, etc. That there were 100 from City α was probably dictated by the method of sampling and that there are 100 in Vocation A, though perhaps not dictated by the method of collecting the data is nevertheless not the issue that concerns us, which is the relative frequencies for the different cities of those in a certain vocation. We shall therefore look upon this particular sample as one of an infinite number of possible samples for each of which the same marginal totals, 100, 200, 300, 100, 90, 140, 75, 85, 110, maintain. The theory that we shall test,—are the relative frequencies in the vocations from city to city such as might be expected due to sampling, is a posteriori to the knowledge of the marginal totals, which is to say that these marginal totals are an intrinsic part of the hypothesis and not issues being tested. Thus the hypothesis being tested is subject to the following restrictions upon the class frequencies:

$$f_{\alpha A} + f_{\alpha B} + f_{\alpha C} + f_{\alpha D} + f_{\alpha E} + f_{\alpha F} + f_{\beta A} + f_{\beta B} + f_{\beta C} + f_{\beta D} + f_{\beta E} + f_{\beta F} + f_{\gamma A} + f_{\gamma B} + f_{\gamma C} + f_{\gamma D} + f_{\gamma E} + f_{\gamma F} = N \quad . \; . \; . \; . \; . \; . \; . \; . \quad [9:12]$$

$$f_{\alpha A} + f_{\alpha B} + f_{\alpha C} + f_{\alpha D} + f_{\alpha E} + f_{\alpha F} = N_{\alpha}$$

$$f_{\beta A} + f_{\beta B} + f_{\beta C} + f_{\beta D} + f_{\beta E} + f_{\beta F} = N_{\beta} \quad . \; . \; . \quad [9:13]$$

$$
\begin{aligned}
f_{\alpha A} + f_{\beta A} + f_{\gamma A} &= N_A \\
f_{\alpha B} + f_{\beta B} + f_{\gamma B} &= N_B \\
f_{\alpha C} + f_{\beta C} + f_{\gamma C} &= N_C \\
f_{\alpha D} + f_{\beta D} + f_{\gamma D} &= N_D \\
f_{\alpha E} + f_{\beta E} + f_{\gamma E} &= N_E
\end{aligned}
\qquad [9{:}14]
$$

There are eight restrictions upon the frequencies in the classes. These restrictions are of three sorts, one is connected with N, the size of the particular sample, the number of the second sort is one less than the number of rows, and the number of the third sort is one less than the number of columns. None of these restrictions can be derived from the others, so there are just eight independent linear restrictions. Such a restric- as

$$f_{\gamma A} + f_{\gamma B} + f_{\gamma C} + f_{\gamma D} + f_{\gamma E} + f_{\gamma F} = N_\gamma$$

is not included for it is derivable from the first three given.

Having thus fixed the marginal totals we know the exact theoretical cell frequencies, under the hypothesis. For any cell, such as that at the intersection of the s row and the t column, this theoretical cell frequency is

$$\tilde{f}_{st} = \frac{f_s}{N} \times \frac{f_s}{N} \times N = p_s p_t N \qquad [9{:}15]$$

Theoretical cell frequency under hypothesis of independence

in which p_s is the proportion in row s and p_t the proportion in column t and N the size of the sample.

An elmentary theorem of probability is that *the probability of the joint occurrence of any number of independent events is the product of the*

probabilities of these events separately. The probability of being in row s is p_s, in column t it is p_t so the probability of a case falling in cell st is $p_s p_t$, and the "expected" number in this cell is $p_s p_t N$, and is frequently designated E_{st}. We shall employ $\tilde{f}_{st}$ for, of course, we should not actually expect that for any cell a sample will have the exact relative frequency of the population.

To reinterpret: The statement in the first sentence of this section, that a priori class frequencies are unknown is false when the hypothesis being tested is so cast as to be contingent upon the observed marginal totals. Under a hypothesis so conditioned we do exactly know the a priori, or theoretical cell frequencies.

Clearly, by-and-large, the observed class frequencies cannot vary as much from the theoretical class frequencies if there are linear restrictions placed upon the class frequencies as if there are no such restrictions. *It has been es-*

TABLE IX D

VOCATIONAL FREQUENCIES FOR DIFFERENT COMMUNITIES

	VOCATIONS						
COMMUNITIES	A	B	C	D	E	F	
α	18. 16.67 1.33 .107	16. 15.00 1.00 .067	22. 23.33 -1.33 .076	13. 12.50 .50 .020	14. 14.17 -.17 .002	17. 18.33 -1.33 .097	$100=N_\alpha$
β	40. 33.33 6.67 1.333	35. 30.00 5.00 .833	44. 46.67 -2.67 .152	20. 25.00 -5.00 1.000	21. 28.33 -7.33 1.898	40. 36.67 3.33 .303	$200=N_\beta$
γ	42. 50.00 -8.00 1.280	39. 45.00 -6.00 .800	74. 70.00 4.00 .229	42. 37.50 4.50 .540	50. 42.50 7.50 1.324	53. 55.00 -2.00 .073	$300=N_\gamma$
	$100=N_A$	$90=N_B$	$140=N_C$	$75=N_D$	$85=N_E$	$110=N_F$	$600=N$

tablished by Fisher (1922) *that a total* χ^2 *from a table having* $(i + g)$ *cells and* g *independent linear restrictions upon the cell frequencies has exactly the same form of distribution as a* χ^2 *from a table having* i *cells and no restrictions upon the cell frequencies.* It is not the number of cells in the table, but i, the number of degrees of freedom, which is crucial in determining the χ^2 distribution.

$$\chi^2 = S\,\chi^2_{st} = 10.134$$

$$\text{d.o.f.} = 10$$

$$\frac{\chi^2}{\text{d.o.f.}} = 1.0134$$

The four entries in a cell are, in order

$$f_{st} = \text{the observed cell frequency} \qquad [9:16]$$

$$\tilde{f}_{st} = \text{the theoretical cell frequency} \qquad [9:17]$$

$$(f_{st} - \tilde{f}_{st}) = \text{the cell divergence} \qquad [9:18]$$

$$\frac{(f_{st} - \tilde{f}_{st})^2}{\tilde{f}_{st}} = \chi^2_{st} = \text{the cell square contingency} \qquad [9:19]$$

For the table entire we have,

$$\chi^2 = S\,\chi^2_{st} \quad \ldots \quad \text{Chi-square, the square contingency} \qquad [9:20]$$

SECTION 5. TRANSFORMATIONS NORMALIZING THE VARIANCE RATIO DISTRIBUTION

V_i ($= \chi_i^2$ as given in [8:38]) is a variance of the sum of i independent standard scores, and V_j another similar variance, independent of the former, having j degrees of freedom.

$$F_{ij} = \frac{V_i/i}{V_j/j} = \text{A variance ratio having } i \text{ d.o.f. in the numerator and } j \text{ d.o.f. in the denominator} \quad [9:21]$$

If the hypothesis is that $V_i/i = V_j/j$, the divergence of F_{ij} from 1.00 (more precisely from its median value) is a measure of improbability of the hypothesis. We obtain a probability measure of any such divergence as follows:

$$\text{Let } f_{ij} = \sqrt[3]{F_{ij}} \quad [9:22]$$

$$\text{Let } \theta_i = \frac{3}{\sqrt{2}} - \frac{\sqrt{2}}{3i}, \text{ and } \theta_j = \frac{3}{\sqrt{2}} - \frac{\sqrt{2}}{3j} \quad [9:23]$$

Table IX E, which is an abridgement of the table of θ_i values given in *The Kelley Statistical Tables*, revision of 1947, facilitates the use of [9:24].

TABLE IX E

TABLE OF θ VALUES

i	θ_i	i	θ_i	i	θ_i	i	θ_i
1	1.6499	17	2.0936	33	2.1070	49	2.1117
2	1.8856	18	2.0951	34	2.1075	50	2.1119
3	1.9642	19	2.0965	35	2.1079	60	2.1135
4	2.0035	20	2.0978	36	2.1082	70	2.1146
5	2.0270	21	2.0989	37	2.1086	80	2.1154
6	2.0428	22	2.0999	38	2.1089	90	2.1161
7	2.0540	23	2.1008	39	2.1092	100	2.1166
8	2.0624	24	2.1017	40	2.1095	200	2.1190
9	2.0689	25	2.1025	41	2.1098	300	2.1197
10	2.0742	26	2.1032	42	2.1101	400	2.1201
11	2.0785	27	2.1039	43	2.1104	500	2.1204
12	2.0820	28	2.1045	44	2.1106	600	2.1205
13	2.0851	29	2.1051	45	2.1108	700	2.1206
14	2.0876	30	2.1056	46	2.1111	800	2.1207
15	2.0899	31	2.1061	47	2.1113	900	2.1208
16	2.0919	32	2.1066	48	2.1115	1000	2.1208
						∞	2.121320

$E^{ii} < .0002$

$$d = \frac{-\theta_i + \theta_j f_{ij}}{\sqrt{\frac{1}{i} + \frac{1}{j} f_{ij}^2}} \qquad \text{The } d \text{ normalizing transformation} \qquad [9:24]$$

Treating d as a deviate in a unit normal distribution will yield q, the proportion to the right of the point d. This is the desired P for F_{ij}, the probability that a variance ratio as great as that observed would arise as a matter of chance under the hypothesis that $\tilde{V}_i/i = \tilde{V}_j/j$. This approximation is close [see Kelley, *Tables* (1947)] if $j > 3$, and if $j < 3$ the xd transformation herewith is close.

$$x = d\left(1 - \frac{.0800}{j^3} d^4\right) \qquad \text{The } xd \text{ normalizing transformation (use when } j < 3) \qquad [9:25]$$

A P from d [9:24], or from x [9:25], having a value in the neighborhood of .50 is most confirmatory of the hypothesis. A small value of P tends to disprove it by asserting that a situation as extreme as the observed situation is very unlikely to arise as a matter of chance and a large value of P tends to disprove it by asserting that as close a fit (agreement between V_i/i and V_j/j) as the observed fit is very unlikely to arise as a matter of chance. We elaborate upon this point in the next paragraphs for it has frequently been misunderstood.

We may write $V_i = V_{i\,(\text{non-chance})} + V_{i\,(\text{chance})}$, that is, the observed variance differs from the underlying true variance by an amount, $V_{i\,(\text{chance})}$, —commonly called the error variance,—whicn is due to sampling. Also $V_j = V_{j\,(\text{non-chance})} + V_{j\,(\text{chance})}$, though in most cases the hypothesis

is such that $V_{j\,(\text{non-chance})} = 0$. Then $V_j = V_{j\,(\text{chance})}$ and the F_{ij} test is to see if $V_{i\,(\text{non-chance})}$ exists. Expressed in terms of the city-vocation distribution problem the test is to see if the relative differences in frequencies of vocations from city to city are such that chance is insufficient to account for them. Thus

$$F_{ij} \text{ becomes } \frac{[V_{i\,(\text{non-chance})} + V_{i\,(\text{chance})}]/\,i}{[V_{j\,(\text{chance})}]\,/\,j}$$

Since $V_{i\,(\text{chance})}/i$ only differs from $V_{j\,(\text{chance})}/j$ as a matter of chance the variance ratio hovers around the value 1.0 when $V_{i\,(\text{non-chance})} = 0$.

If $V_{i(\text{non-chance})}$ is substantial and

$$\frac{V_{i\,(\text{chance})}}{i} = \frac{V_{j\,(\text{chance})}}{j}$$

then F_{ij} *is much greater than* 1.0 and P is small. This situation is easy to interpret. We simply believe the hypothesis disproved because of the presence in V_i of non-chance factors.

If F_{ij} is much less than 1.0 we not only must believe that $V_{i\,(\text{non-chance})}$ is nonexistent or negligible, but also that $V_{i\,(\text{chance})}$ is unreasonably small, again disproving the hypothesis. This situation is the more difficult to account for. It can be brought about (as certain poorly devised studies bear witness to) if an experimental procedure has employed such controls as to eliminate to a degree the operation of the same chance factors in V_i as operate in V_j.

In short, if F_{ij} is large and P small consider the hypothesis untenable because of variability factors in V_i in excess of the hypothesis, and if F_{ij} is small and P large consider it untenable

because of faulty controls in the experimental procedure. This statement of the case holds when the denominator variance, V_j, is the chance variance. The writer recommends that F_{ij} be so written that this is always the case no matter if this result in an F_{ij} which is less than 1.0. In order to use certain tables it has been customary so to write F that it is greater than 1.0. It is only necessary to note, if P_{ij} is P from F_{ij} and P_{ji} is P from F_{ji},

$$(F_{ji} = \frac{1}{F_{ij}})$$

that

$$P_{ij} = 1 - P_{ji} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; [9:26]$$

to see that no matter how it is written to conform to tables it can always be written with the error variance in the denominator for purposes of interpretation.

Since χ^2 / d.o.f. from Table IX D is equal to V_i/i in which i = 10, we have

$$F_{10,\infty} = \frac{10.134}{10.} = 1.0134$$

Since $F_{10,\infty}$ so nearly equals 1.0 we can, without computing P, be convinced that cell-divergences from theoretical values as great as the observed divergences could readily arise as a matter of chance. The determination of P merely confirms this. We have

$F_{10,\infty}$	= 1.0134	θ_∞'	= 2.121320
$f_{10,\infty}$	= 1.00445	$\sqrt{.1}$	= .316228
		d	= .17885
θ_{10}	= 2.0742	P	= .4290*

* P computed by linear interpolation in Pearson's TABLES OF χ^2 (1914) is .4294.

Thus there are 43 chances in 100 that if $(V_1/i) = 1.0$ a sample value as large as 1.0134 would occur.

We find the data highly consistent with the hypothesis tested, but we should not say that the hypothesis has been "proven", for the data may be consistent with any number of other hypotheses. *Though a high or a low P can be taken as "disproof" of the hypothesis, one in the neighborhood of .5 is only indicative of consistency with the hypothesis, not of its proof.* We never prove a hypothesis in the sense that it and none other is tenable, but only at times have evidence which does not disprove it.

The accuracy of P from the d, or the xd-transformation extends to the third or fourth decimal places and is sufficient for all experimental purposes. A more serious hazard is consequent to the assumption that $(f_{st}-\tilde{f}_{st})$ is normally distributed. The actual divergence from normality is due to the fact that f_{st} takes integral values only, and further, if p_{st} and N are both small the distribution of $(f_{st}-\tilde{f}_{st})$ approaches a Poisson and not a normal distribution. To lessen this hazard Kelley (1923) recommended that $(p_{st}N)$ be > 1; Fisher and Yates (1938) recommend such a grouping that $(p_{st}N)$ be not < 5.0; and Cochran (1942) provides data upon the requisite minimal size of $(p_{st}N)$ at the one and the five per cent levels of significance for different degrees of freedom. He writes, "With only a single small expectation [i.e., only one cell with small $(p_{st}N)$] the conventional limit of five for the smallest expectation appears unduly high. At the 5 per cent level, the tabular χ^2 distribution may be used without undue error with an expectation as low as 0.5, and at the one per cent level with an expectation as low as 2." With one degree of freedom Yates' correction is to be applied, though

Cochran points our certain infrequent exceptions to this. A correction of the Yates type is in order when the number of degrees of freedom exceeds one, but Cochran has shown that its value is small and its precise computation rather complicated.

A fundamental property of χ^2's is given by the following equation, in which the subscripts indicate the respective degrees of freedom:

$$\chi_i^2 + \chi_j^2 + \ldots + \chi_t^2 = \chi^2_{i+j+\ldots+t} \qquad \text{Additive property of } \chi^2 \qquad [9:27]$$

p from $\chi^2_{i+j+\ldots+t}$ is the probability that chance would yield divergences-squared in all the situations whose sum total is as great as that observed. *This frequently enables the combination of a number of experiments so as to obtain a single measure of confidence. However, if the sign of a divergence is material, the method does not answer the question of importance.* What is then required is a method based upon deviations with their sundry plus and minus signs and not one based upon deviations squared.

χ^2 has been extensively used in such contingency situations as the City-Vocation problem here given as an illustration, but theoretically it is more exactly applicable to quantitative problems wherein the variable is both continuous and more nearly normal than is the frequency in a class variable. Generally, where quantitative data are involved, more informative regression, analysis of variance, and variance ratio (F_{ij}, not merely $F_{i\infty}$ which = χ^2/d.o.f.) methods are available, as illustrated in Chapter X. The student should not believe that the χ^2 method is in any way limited to contingency data.

CHAPTER X

ESTIMATION, REGRESSION, AND CORRELATION

SECTION 1. A BRIEF PERSPECTIVE OF THE FIELD

It has been claimed that every properly conducted experiment should hinge upon some null hypothesis. Most correlation problems are not so stated. If a new psychological test is being developed for use in connection with college admission, the initial null hypothesis that the test correlates with college success to the extent zero is not made. If the correlation found is .70 (st. er. = .03) something very useful is now known that was not known before, but there has been no null hypothesis. We shall here be concerned with correlation and regression methods without regard to null hypotheses, but will later note how nicely they fit in with certain null hypotheses and how gracefully they contribute to the analysis of variance.

The correlation coefficient, r_{01}, is a measure of mutual or reciprocal relationship and is frequently informative of itself, though its primary utility is simple as an essential measure entering into an estimation equation.

$$\bar{X}_0 = a_{01} + b_{01}X_1 \qquad \text{Linear regression equation of } X_0 \text{ upon } X_1 \qquad [10:01]$$

$$a_{01} = M_0 - b_{01}M_1 \quad \ldots\ldots\ldots\ldots\ldots \quad [10:02]$$

$$b_{01} = r_{01}\frac{\sigma_0}{\sigma_1} \qquad \text{Regression coefficient of } X_0 \text{ upon } X_1 \qquad [10:03]$$

Thus

$$\bar{X}_0 = M_0 - r_{01}\frac{\sigma_0}{\sigma_1}M_1 + r_{01}\frac{\sigma_0}{\sigma_1}X_1 \quad \ldots\ldots \quad [10:01a]$$

When deviation scores are employed [10:01] becomes

$$\bar{x}_0 = b_{01}x_1 = r_{01}\frac{\sigma_0}{\sigma_1}x_1 \quad \ldots\ldots\ldots \quad [10:04]$$

and by parity

$$\bar{x}_1 = b_{10}x_0 = r_{01}\frac{\sigma_1}{\sigma_0}x_0 \quad \ldots\ldots\ldots \quad [10:04a]$$

When standard scores (see [8:23]) are employed [10:01] becomes

$$\bar{z}_0 = r_{01}z_1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad [10:05]$$

and by parity

$$\bar{z}_1 = r_{01}z_0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad [10:05a]$$

revealing the reciprocal nature of the relationship.

There are many correlation formulas because of the variety of types of data and because of the need for various corrections in the raw results. We will designate the four most important corrections by preceding subscripts: *a* for attenuation, *c* for coarseness of grouping, *f* for fineness of grouping, and *s* for shrinkage. Which of these corrections applies varies from situation to situation because of the differences in types of data and of uses to which results may

be put. The corrections listed in Table X A have been developed in connection with quantitative data, and the fineness of grouping correction is of prime importance in connection with qualitative data.

TABLE X A

NOTATION OF SUNDRY CORRECTIONS TO CORRELATION COEFFICIENTS

r_{01},	raw r between quantitative variables X_0 and X_1
${}_{a_0a_1}r_{01}$,	also designated $r_{\infty\omega}$, is r_{01} corrected for attenuating effect of errors of measurement of X_0 and X_1
${}_{a_0}r_{01}$,	also designated $r_{\infty 1}$, is r_{01} corrected for attenuating effect of errors of measurement of X_0
${}_{a_1}r_{01}$,	also designated $r_{0\omega}$, is r_{01} corrected for attenuating effect of errors of measurement of X_1
${}_cr_{01} \equiv {}_{c_0c_1}r_{01}$,	r_{01} corrected for equally spaced coarse grouping of X_0 and X_1
${}_{c_0}r_{01}$,	r_{01} corrected for equally spaced coarse grouping of X_0
${}_{c_1}r_{01}$,	r_{01} corrected for equally spaced coarse grouping of X_1
${}_{c'}r_{01} \equiv {}_{c_0'c_1'}r_{01}$,	r_{01} corrected for unequally spaced coarse grouping of X_0 and X_1
${}_sr^2_{01}$,	r^2_{01} corrected for shrinkage

r, or r^2, with multiple preceding subscripts are combinations of the preceding

In case regression is nonlinear one of the variables, usually X_1, may be transformed into a new variable, X_1', producing substantial linearity and the correlation $r_{01'}$, computed. This transformation may frequently be accomplished by a higher order parabolic regression. If of the second degree, the situation is equivalent to a multiple variable problem, involving three variables, X_0, X_1, X_1^2, the multiple correlation notation for which is $r_{0\Delta 1,1^2}$

$$\overline{\overline{X}}_0 = a + b\,X_1 + c\,X_1^2 \qquad \text{Second degree parabolic regression of } X_0 \text{ on } X_1 \qquad [10:06]$$

The double bar is used to distinguish this estimated X_0 from that given by [10:01]. The correlation between X_0 and the right-hand member of [10:06] is $r_{0\Delta 1,1^2}$ and it may call for a, c, and s corrections.

This variety of corrections would be very troublesome were it not ordinarily possible to work with such original data as not to demand them. If N, the number of cases in the sample, is large the shrinkage correction is negligible. The grouping of measures into classes is usually upon the basis of equally spaced intervals so the c' correction is not involved. Usually 12 or more equally spaced classes can be employed, thus making the c correction unimportant. The "correction for attenuation" adjustment yields an estimate of what the correlation would be were errors of measurement lacking. It thus answers a theoretical question and is not appropriate when the problem is to estimate an actual X_0 (not a hypothetical one without error) from an actual X_1 (not from a hypothetical one without error).

Though, for a particular situation, these corrections may not be demanded, it is only after the student has thought of them in turn and decided that they are inappropriate (as the a cor-

rection), or small (as the c, s, or nonlinear adjustment) for his problem that he should refrain from making them, if it is possible to make them. Not infrequently the logic of the problem calls for a correction for attenuation but the requisite data for making the correction are lacking. In such instance a definitely expressed reservation in interpretation is called for.

An important classification of correlation procedures depends upon the type of variable involved, as listed in Table X B.

TABLE X B

TYPES OF AVAILABLE VARIABLES

(a) Two-category variable
  (a-1) Values important as given, e.g., male and female.
  (a-2) Values are crude measures of an underlying continuous trait, e.g., above-average and below-average intelligence.

(b) Categorical variable, consisting of several values which are not known to be quantitatively related, e.g., nationalities.

(c) Ordered variable, consisting of several values which may be placed in an order of magnitude, but not quantitatively expressed, e.g., a number of people ranked in order for sociability.

(d) Quantitative variable, e.g., height.

The field is fairly adequately covered by Tables X A and X B, but neither is inclusive of all the possible situations that should be distinguished because of the statistical consequences that follow. Even so, when the various corrections are combined with the various types of variables there result over a hundred easily distinguishable and definable correlation situations.

Of all these situations the most fundamental one is that which requires no *a* correction because the variables are fully meaningful as they stand, no *c* correction because each variable is recorded in 12 or more equally spaced intervals, no *s* correction because the size of the sample is 25 or greater, and no adjustment for nonlinearity.

The data of Francis Galton upon the relationship of height of offspring and of parents, which meet all these conditions (though Galton used fewer classes than 12), will be used in the next two Sections for illustrative purposes. At the same time, a noting of Galton's study provides a striking illustration of scientific induction.

SECTION 2. THE PROBLEM OF CONCOMITANT VARIATION IN THE SCIENCES

A problem common to all sciences is that of defining and determining the cause of concomitant variation. In this the physical sciences have a great advantage over the biological and social sciences in that (1) errors of observation and measurement are usually very small in comparison with the measurements involved and (2) fewer factors are ordinarily present. In measuring some intellectual capacity of a group of children, it usually happens that the standard errors of the test scores obtained are greater than half the standard deviation of the scores of the group. Obviously any relationship between two capacities, each measured with no greater reliability than this, will be clouded by the errors of measurement. This is serious, but it is not the only difficulty. In measuring the effect of gravity, physicists can ordinarily assume that 10 pounds of lead and 10 pounds of iron will act in a similar manner, but in measuring intellect, food prices, etc., to say that one reagent, one commodity, etc., is equivalent to another with re-

spect to the function being examined, is usually questionable. Accordingly, whereas the investigations of physics lead to the establishment of "laws", those of the social sciences ordinarily lead to the discovery of "tendencies". Relationships between two psychological, biological, or social factors frequently depend upon a number of causes, each more or less independent, and no one of such importance as to dominate the situation. Under these conditions, the relationship tends to be linear. In other cases, where the true relationship is nonlinear, large errors of measurement will lessen the strength of the measurable relationship, thereby making it more difficult to determine the exact nature of whatever curvilinear relationship may exist. It is also true that relationships which are intrinsically curvilinear when determined over a range of the two variables from very low to very high, may show practically linear relationship throughout a short stretch of the range. For all the reasons stated, a measure of relationship based upon the assumption of linearity is of great importance. Even in the case of known nonlinear relationship it is of much value as a point of departure. The equation for estimating one variable from a knowledge of another linearly related to it is called a linear regression equation [10:01], and in general the best measure of mutual linear relationship is the Pearson product-moment correlation coefficient.

Fundamental properties of this measure of relationship were discovered and presented graphically by Francis Galton (1877 to 1888). Galton's investigations had to do with the inheritance of traits, and certain of the terms which he used would hardly have arisen if the development had involved other data. For example, the symbol "*r*" was a measure of "reversion", such, for example,

as offspring upon mid-parent (a mid-parent measure is an average of the measures of father and mother). Later, Galton used the terms "regression" and "co-relation" and called the measure the "index of co-relation." Weldon very properly calls this measure "Galton's function" and Edgworth, in 1892, gave it the name which has survived, "coefficient of correlation." Pearson (1920) has pointed out that the product-moment function of Bravais bears but a resemblance in form to the product-moment coefficient of correlation. Whereas Bravais started with observations which were assumed to be independent and obtained derived measures whose product moments did not equal zero, Galton started with the epic-making concept that the original measures were dependent. Partial correlation analysis leads to independent measures, having given related original measures, which is exactly the reverse of the Bravais or Gaussian developments. Galton seems deserving of being called the father of correlation, though so ubiquitous a phenomenon as concomitant variation in the natural and social sciences did not completely escape his predecessors and it has been rediscovered by many modern students working in fields remote from those in which Galton labored.

SECTION 3. FINDINGS RESULTING FROM GALTON'S GRAPHIC TREATMENT

Galton's procedure, based upon medians and quartile deviations, has given way to the more accurate one involving the product-moment formula based upon deviations from means and standard deviations, as developed by Karl Pearson.

$$r \equiv r_{12} = \frac{\Sigma x_1 x_2}{N\sigma_1\sigma_2} = \frac{\Sigma xy}{N\sigma_1\sigma_2} = \frac{\Sigma X_1 X_2 - NM_1M_2}{\sqrt{\Sigma X_1^2 - NM_1^2}\sqrt{\Sigma X_2^2 - NM_2^2}} \qquad [10{:}07]$$

Equivalent formulas for Pearson product-moment correlation coefficient (See also (10:27) and (10:29)

We shall use Galton's data in deriving a measure of correlation. Galton obtained the heights of parents and the heights of children, and drew up a "correlation" table or "scatter diagram" showing the relationship between the two. All female heights were multiplied by 1.08 to make them comparable with male heights. This is not the soundest procedure, but in this problem leads to no material error. Letting X_1, X_2 represent male and female heights, σ_1, σ_2 their standard deviations, and M_1, M_2 their means, it would have been better to have reduced each female height to a comparable male height by the equation

$$\text{Comparable male height} = M_1 + (X_2 - M_2)\frac{\sigma_1}{\sigma_2} \qquad [10:08]$$

The discussion which follows will assume that

CHART X I

CORRELATION BETWEEN HEIGHTS OF MIDPARENTS AND OFFSPRING

HEIGHTS OF ADULT CHILDREN EXPRESSED AS DEVIATIONS FROM THE MEAN HEIGHT 68¼ INCHES

X_2

HEIGHTS OF MIDPARENTS X_1

	-4½	-3½	-2½	-1½	-½	½	1½	2½	3½	4½	f	ξ	fξ	fξ²	Sfζ	Sfξζ
3½						1	2	2	2	1	8	7	56	392	40	280
2½				2	4	5	5	4	3	1	24	5	120	600	60	300
1½	1	2	3	5	8	9	9	8	5	3	53	3	159	477	77	231
½	2	3	6	10	12	12	12	10	6	3	76	1	76	76	56	56
-½	3	7	11	13	14	13	10	7	3	1	82	-1	-82 411	82	-76	76
-1½	3	6	8	11	11	8	6	3	1		57	-3	-171	513	-105	315
-2½	2	3	4	6	4	3	2				24	-5	-120 -373	600	-72	360
f	11	21	32	47	53	51	46	34	20	9	324 = N		38 $=\Sigma\xi$	2740 $=\Sigma\xi^2$		1618 $=\Sigma\xi\zeta$
ζ	-9	-7	-5	-3	-1	1	3	5	7	9						
fζ	-99	-147	-160	-141	-53 ∫600	51	138	170	140	81 ∫580	∫-20 $=\Sigma\zeta$					
fζ²	891	1029	800	423	53	51	414	850	980	729	∫6320 $=\Sigma\zeta^2$					
Sfξ	-17	-31	-40	-41	-11	19	40	52	44	23						
Sfξζ	153	217	200	123	11	19	120	260	308	207	∫1618 $=\Sigma\xi\zeta$					

the more reliable method of transforming ("transmuting" according to Galton) female into male heights was followed and also that means were used throughout. Presumably Galton used medians, but no fundamental difference in treatment followed from such use, it simply being a slightly less reliable procedure. Galton's diagram contained the data given in the accompanying correlation table or scatter diagram, Chart X I. Deviations being measured from 68.25 inches, which is a small fraction of an inch away from the actual means, are labeled ξ and ζ instead of x and y, and account of this slight difference is taken in Section 5. From just such data as given,—in fact, it is likely that these identical data were involved,—Galton inferred certain relationships which we now know hold with every normal correlation surface [formula 10:10].

(a) A plot of the means of the horizontal arrays (rows) as shown by the x's shows the "reversion" of offspring upon height of mid-parent. Thus if the mid-parent height is 2.50 above the mean, the average or most probable height of offspring is 1.25 inches above the mean.

(b) The line connecting these means may be closely represented by a straight line through the origin, or intersection, of the means of the two distributions. This is the line showing the regression (or "reversion") of offspring upon mid-parent.

(c) There is a reversion or regression of mid-parent upon offspring. This would be represented by a straight line passing approximately through the o's. Thus for every correlation table there are two regression lines.

(d) The slopes of these two lines are equal, provided the standard deviations of the two distributions are equal.

(e) If the standard deviations are equal this slope varies between zero and one (Galton did

not suggest the existence of negative correlations), and may be represented by the symbol "*r*."

(*f*) The standard deviations of the measures found in successive arrays (successive rows, or successive columns) are approximately equal and are smaller than the standard deviations of the total distribution, so that, if σ_2 equals the standard deviation of the heights of offspring and $\sigma_{2.1}$ equals the standard deviation of offspring corresponding to given heights of mid-parent, then

$$\sigma^2_{2.1} = \sigma^2_2\ (1 - \lambda)$$

where λ is a positive quantity less than 1.0. Also, dealing with columns instead of rows,

$$\sigma^2_{1.2} = \sigma^2_1\ (1 - \lambda)$$

in which λ is the same as before, σ_1 the standard deviation of heights of mid-parents, and $\sigma_{1.2}$ the standard deviation of heights of mid-parents corresponding to given heights of offspring. These relationships, which Galton discovered in his data, will require re-examination and restatement (see Chapter XI, Section 4) for less linear and less homoscedastic (less equal-variability of arrays) data.

(*g*) There is a simple relationship between λ and *r*. It is

$$\lambda = r^2, \text{ so that}$$

$$\sigma^2_{2.1} = \sigma^2_2\ (1-r^2) \qquad \sigma^2_{1.2} = \sigma^2_1\ (1-r^2) \qquad [10:09]$$

Variance of deviations from points on regression line (see [10:24] and [10:49])

(*h*) Each array is approximately normal if the

total distributions are normal.

(i) If contour lines for different frequency densities are drawn in the diagram, they constitute a system of similar and similarly placed ellipses, the conjugate diameters of which are the two regression lines.

Galton made no claim to mathematical ability, but through sheer insight into the phenomena of mutual implication made these penetrating observations. He carried his conclusions, stated in probability terms, as to the nature of the correlation surface, to J. D. Hamilton Dickson (1886), a mathematician, who readily wrote down the normal correlation equation involving two variables. In the notation here used this is:

$$y = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \frac{-1}{2(1-r^2)} \left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} - \frac{2x_1x_2r}{\sigma_1\sigma_2}\right) \qquad [10:10]$$

Normal correlation surface, two variables

Galton's humility, after years of collection of data and subtle analysis of the same, in the face of the neat but not involved mathematical derivation, is worthy of note by the social scientists of this day who scoff at mathematical analysis. Upon receiving the solution of his problem from Mr. Dickson, he wrote:* "I may be permitted to say that I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis as when this answer reached me, confirming, by purely mathematical reasoning, my various and laborious statistical conclusions with far more minuteness than I had dared to hope, for the original data ran somewhat roughly, and I had to smooth them with tender caution."

* See Karl Pearson, NOTES, 1920.

SECTION 4. ALGEBRAIC STATEMENT OF GALTON'S GRAPHIC FINDINGS AND DERIVATION OF CORRELATION FORMULAS

Let us consider these discoveries in detail. Let x_1 stand for the height of mid-parent, x_2 for that of offspring, each as a deviation from its mean. Let the variances be V_1 and V_2 and the standard deviations σ_1 and σ_2, while r is the slope of the regression line in the "reduced" scatter diagram, i.e., in a correlation table in which the measures entered are x_1/σ_1 and x_2/σ_2. Galton reduced by dividing by the quartile deviations, leading to essentially the same result as here. The slopes of the two regression lines are equal and equal to r. We will shortly obtain a value of r by means other than the graphic method of Galton. Finally, let $\bar{x}_2$ stand for an estimated height of offspring, knowing the mid-parent height and let $\bar{x}_1$ be an estimated height of mid-parent, knowing the offspring height. An alternative and more precise notation for $\bar{x}_2$ is $x_{2\Delta 1}$, which may be read "x_2 dependent upon x_1." Similarly, an alternative notation for $\bar{x}_1$ is $x_{1\Delta 2}$. With this notation, discoveries *(a)* and *(b)* together are equivalent to

$$\frac{\bar{x}_1}{\sigma_1} = r\frac{x_2}{\sigma_2}$$

Fundamental form of regression equation [see 10:05]

Propositions *(c)* and *(d)* are equivalent to the addition of the second fundamental regression equation, as follows:

$$\frac{\bar{x}_2}{\sigma_2} = r\frac{x_1}{\sigma_1}.$$

Proposition *(e)* is liable of misinterpretation. If $r=0$, it asserts that there is no relationship

between the variables,—no regression of one variable upon the other,—while $r=1$ means complete mutual implication. More loosely stated, this latter situation will be described as one of complete dependence. So far as the data are concerned there is no evidence that the heights of parents have any more to do in causing the heights of offspring than do the heights of offspring in causing the heights of the parents. This is still true should $r=1$. A situation exists and a correlation coefficient measures the tendency of the paired observations to be related, but it gives no evidence whether x_1 is the cause of x_2, x_2 the cause of x_1, or whether the cause is unknown and lies back of both. We think of parents being causal agents in determining the heights of offspring, but we do this for reasons outside of the scatter diagram, namely, the parents have existed earlier than the offspring in a time series.

Propositions *(f)* and *(g)* are the result of careful study of data, but Galton gave a simple proof of *(g)*. The variability of the offspring generation is consequent to the variability of the arrays (rows) and the variability of the means of these arrays. If $x_{2\Delta 1_i}$ is the distance of the mean of the *i*-array of offspring heights from the mean of all offspring heights and, as before, $V_{2.1}$ is the variance of this as well as of each of the other arrays, and if n_1 is the number of measures in this array, then $(n_i V_{2.1} + n_i x^2_{2\Delta 1_i})$ is the contribution of the *i*-array in the calculation of the variance of the distribution, thus:

$$V_2 = \frac{\sum^k n_i V_{2.1}}{N} + \frac{\sum^k n_i x^2_{2\Delta 1_i}}{N} \qquad \text{Analysis of } V_2 \text{ variance} \qquad [10:11]$$

in which k over the summation sign indicates the number of terms in the summation, i.e., k is the number of classes into which the x_1 variable is grouped. The first term of the right-hand member yields

$$\frac{\sum^{k} n_i V_{2.1}}{N} = V_{2.1} \frac{\sum^{k} n_i}{N} = V_{2.1} \quad . \; . \; . \; . \; . \quad [10:12]$$

Since, for any array,

$$x_{2\Delta 1_i} = r \frac{\sigma_2}{\sigma_1} x_{1_i} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad [10:13]$$

and since

$$\sum^{k} n_i \; x^2_{1_i} \equiv \sum^{n} x_1^2 = N \; V_1 \quad . \; . \; . \; . \; . \; . \; . \quad [10:14]$$

we obtain for the second right-hand term of [10:11] $V_{2\Delta 1}$, the variance of the points on the regression line

$$V_{2\Delta 1} = \frac{\sum^{k} n_i x^2_{2\Delta 1_i}}{N} = r^2 V_2 \qquad \text{Variance of estimated } x_2\text{'s} \quad [10:15]$$

It is important to distinguish between this, the variance of the points on the regression line, and [10:50], the variance error of a point on the regression line.

$V_{2\Delta 1}$ is read as "the variance of that part of x_2 that is dependent upon x_1."

Accordingly, we have

$$V_2 = V_{2.1} + V_{2\Delta 1} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [10:16]$$

Total variance in terms of the variance of arrays and of the variance of points on linear regression line

$$V_2 = V_{2.1} + r^2 V_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . [10:17]$$

Transposing terms

$$V_{2.1} = V_2(1-r^2) \quad \text{Variance of an array} \quad [10:18]$$

Also

$$V_{1.2} = V_1(1-r^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . [10:18a]$$

This derivation proves Galton's proposition *(g)*, that the reducing factor is equal to r^2.

(h) is an experimental finding which, coupled with *(g)* and *(a)*, *(b)*, *(c)*, and *(d)*, quickly gives the equation of the normal bivariate correlation surface.

A normal equation, with area 1.0, is given herewith [10:19]

$$y = \frac{1}{\sigma \sqrt{2\pi}} \exp \frac{-x^2}{2V} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [10:19]$$

Let any integral value of x include all cases with values between $(x-.5)$ and $(x+.5)$. Further, y is the ordinate at the point x, so that $y \times 1$ is the area of a rectangle with ordinate y and base from $(x-.5)$ to $(x+.5)$. Thus, if the (interval/σ) is small,—and let us here employ such units of measurement that this is so,—then this rectangle closely approximates the curve for the region from $(x-.5)$ to $(x+.5)$; and y is the probability that a case drawn at random will have the value x.

Referring to the two-dimensional scatter diagram, a similar statement for the probability that a measure in the x_2 array shall have the value x_1 is $y_{1.2}$,—the ordinate for variable values of x_1 for a fixed value of x_2.

$$y_{1.2}=\frac{1}{\sigma_1\sqrt{1-r^2}\sqrt{2\pi}}\exp\frac{-(x_1-r\frac{\sigma_1}{\sigma_2}x_2)^2}{2V_1(1-r^2)} \qquad [10:20]$$

The probability that a measure will lie in the x_2 array is

$$y_2=\frac{1}{\sigma_2\sqrt{2\pi}}\exp\frac{-x_2^2}{2V_2} \qquad [10:19]$$

The probability of the joint occurrence, or that a case chosen at random will have the value x_2 and the value x_1 is the product of these two probabilities, or $y = y_2 y_{1.2}$, which reduces to

$$y=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\exp\frac{-1}{2(1-r^2)}\left\{\frac{x_1^2}{V_1}-2\frac{x_1}{\sigma_1}\frac{x_2}{\sigma_2}r+\frac{x_2^2}{V_2}\right\}$$

Normal bivariate distribution of volume 1.0 [10:21]

Multiplying the right-hand member by N yields the normal bivariate distribution with number of cases equal to N, which Dickson wrote down to satisfy the experimental findings that Galton reported to him.

SECTION 5. DERIVATION OF COMPUTATIONAL FORMULAS FOR b AND r

The magnitude r has, to this point, been defined as the slope of the regression line in case standard scores (see [10:05] and [10:05a]) are the measures entered into the correlation table. We will now prove that in any scatter diagram, the two "best fit" linear regression lines are, as recorded, [10:04] and [10:04a], in which r is readily computable by formula [10:29] and the b's

by formulas [10:30] and [10:31].

The term "best fit" is used as in the method of least squares. A "best fit" determination is one in which the variance error of estimate is minimal. Determinations can be made resulting in the sum of the deviations, of the cubes of the deviations, of their fourth powers, etc., being a minimum, but since the days of Gauss, it has been known that in the case of a normal distribution none of these determinations will result in as small a median error as one in which the sum of the squares of the errors of estimate is made a minimum. The constants of distributions which are quite divergent from the normal, so determined that the variance error is minimal, are undoubtedly excellent determinations, but it is no longer possible to say that constants so calculated have smaller median errors than would others derived upon a different principle. In derivations herewith the Gaussian principle of least squares is adhered to.

In Chart X I the regression line drawn is that of x_2 upon x_1. Its equation is

$$\bar{x}_2 = r\frac{\sigma_2}{\sigma_1}x_1 = b_{21}\,x_1 \qquad [10:22]$$

The "slope" of the regression line is b_{21} and equals tan ϕ. Having given a value of x_1, the best estimate of the corresponding x_2 is $\bar{x}_2$,— [10:22]. In general x_2 will not be identical with the actual or experimentally obtained value of x_2, so that

$$x_{2.1} = x_2 - \bar{x}_2 \quad \text{Error of estimate} \qquad [10:23]$$

is an error of estimate. The variance error of estimate is given by

$$\sigma^2_{2.1} = V_{2.1} = \frac{\Sigma(x_2 - \bar{x}_2)^2}{N}$$ Variance error of estimate (see [10:09] and [10:49]) [10:24]

The value of b_{21} which makes this minimal is the value sought.

$$N\ V_{2.1} = \Sigma(x_2 - \bar{x}_2)^2 = \Sigma(x_2 - b_{21}x_1)^2$$

$$= \Sigma x_2^2 - 2b_{21}\ \Sigma x_1 x_2 + b_{21}^2\ \Sigma x_1^2$$

$$= N\ V_2 - 2\ N\ b_{21}\ c_{12} + N\ b_{21}^2\ V_1$$

in which c_{12}, called a covariance, or product moment, equals $(\Sigma x_1 x_2)/N$. Dividing by N and completing the square on the last two terms, we have

$$V_{2.1} = V_2 + \frac{c_{12}^2}{V_1} - 2\ b_{21}\ c_{12} + b_{21}^2 V_1 - \frac{c_{12}^2}{V_1}$$

$$= V_2 - \frac{c_{12}^2}{V_1} + (\frac{c_{12}}{\sigma_1} - b_{21}\ \sigma_1)^2$$

As the () term is squared, it can never be negative, so obviously $V_{2.1}$ takes its proper, or smallest, value when the () term, which is the only term containing b_{21}, equals zero. Solving

$$b_{21} = \frac{c_{12}}{V_1} = \frac{\Sigma x_1 x_2}{N\ V_1} = \frac{\Sigma x_1 x_2}{\Sigma x_1^2}$$ Regression coefficient of x_2 upon x_1 [10:25]

By parity

$$b_{12} = \frac{c_{12}}{V_2} = \frac{\Sigma x_1 x_2}{N\ V_2} = \frac{\Sigma x_1 x_2}{\Sigma x_2^2}$$ [10:26]

Regression coefficient of x_1 upon x_2

The correlation coefficient is either regression coefficient, when standard scores are involved, since they are then equal. To show that this common value is r as given by [10:07] is left as an exercise.

When $b_{21} \neq b_{12}$ we can note that the sign of both b's is that of $\Sigma x_1 x_2$, so that

$$r_{12} = \sqrt{b_{12}\ b_{21}} \quad \ldots\ldots\ldots \quad [10:27]$$

with the sign attached to r that is the sign of b_{12} or of b_{21}.

The formulas for r and b based upon deviation scores, x_1 and x_2, are unsatisfactory for computational purposes because x_1 and x_2 would ordinarily be continuing decimals. We require a method utilizing integral values corresponding to the midpoints of the classes into which the X_1 and X_2 scores have been grouped. To avoid a notation calling for subscripts of subscripts, we let X, x, and ξ have meanings for the X_1 variable as already defined [Ch. VI, Sec. 6]. The grouping interval in this first variable is i_1. We let Y, y, ζ, and i_2 have corresponding meanings for the X_2 variable. Formulas for the evaluation of M_1, V_1, M_2, and V_2 in terms of ξ and ζ scores and of i_1 and i_2 have already been given, so we only need to express c_{12} in these terms to have available computational procedures based upon the convenient ξ and ζ scores.

$$N\ c_{12} = \Sigma xy = \Sigma[(\xi - M_\xi)i_1][(\zeta - M_\zeta)i_2]$$

$$= i_1 i_2(\Sigma\xi\zeta - M_\xi\Sigma\zeta - M_\zeta\Sigma\xi + N\ M_\xi M_\zeta)$$

$$c_{12} = \frac{i_1 i_2}{N}(\Sigma\xi\zeta - N\ M_\xi M_\zeta) \qquad [10:28]$$

Covariance from ξ and ζ scores

Utilizing this with earlier formulas we have

$$r_{12} = \frac{c_{12}}{\sigma_1\sigma_2} = \frac{\Sigma\xi\zeta - N\ M_\xi M_\zeta}{N\ \sigma_\xi\ \sigma_\zeta} \quad \text{Work formula for } r \qquad [10:29]$$

$$b_{12} = \frac{c_{12}}{V_2} = \frac{i_1(\Sigma\xi\zeta - N\ M_\xi M_\zeta)}{i_2\ N\ V} \quad \text{Work formula for } b_{12} \qquad [10:30]$$

$$b_{21} = \frac{c_{12}}{V_1} = \frac{i_2(\Sigma\xi\zeta - N\ M_\xi M_\zeta)}{i_1\ N\ V_\xi} \quad \text{Work formula for } b_{21} \qquad [10:31]$$

Computing the correlation and regression constants for the Galton data by means of these formulas we have

$N = 324$

Arbitrary origin$_1$ = 68.25, and $i_1 = .5$

Arbitrary origin$_2$ = 68.25, and $i_2 = .5$

$M_\xi = \frac{38}{324} = .1173$, so that $M_1 = 68.25 + .5(.1173) = 68.31$

$M_\zeta = \frac{-20}{324} = -.0617$, so that $M_2 = 68.25 + .5(-.0617) = 68.22$

$V_\xi = \frac{2740}{324} - (.1173)^2 = 8.4430$, so $V_1 = 8.4430\ (.5)^2 = 2.1108$

$V_\zeta = \frac{6320}{324} - (-.0617)^2 = 19.5024$, so $V_2 = 4.8765$ and $\sigma_2 = 2.2081$

$c_{12} = \frac{(.5)(.5)}{324}[1618 - 324(.1173)(-.0617)] = 1.2503$

$$b_{12} = \frac{1.2503}{4.8765} = .2564$$

$$b_{21} = \frac{1.2503}{2.1108} = .5923$$

$$r_{12} = \frac{1.2503}{(1.4528)(2.2081)} = .3897$$

$$\overline{X}_2 = 27.76 + .5923\ X_1 \quad \text{(Substituting in [10:01a])}$$

The reliability and, accordingly, number of decimal places to be retained, of these correlation constants is discussed in this chapter, Sections 6 and 7.

The computation shown on Chart X I is straightforward, but it does not provide as adequate checks upon numerical accuracy as are desirable. Many forms have been devised to make the steps involved mechanical and such that nonstatistically trained clerks can compute correlational constants rapidly and accurately. Some of these are especially designed for use with computing machines. The form given in Appendix C is an example of one that is serviceable either with or without a computing machine. It is a little longer than some of the forms, but it provides more complete checks than usual for summations are obtained in two independent ways. The steps to be followed will not be explained in detail as they are practically self-explanatory, but attention is called to the following facts: that $s = \xi + \zeta$; that $d = \xi - \zeta$; that the italic entries are the computational or manual entries; that "LL-UR diag" indicates that cells in a line from the lower-left to the upper-right all have the same d-values, so that summations of fre-

quencies along such a line give all the frequencies having the particular d value recorded at the end of this line at the bottom or top of the scatter diagram; that cells in a line from the upper-left to the lower right, "UL-LR diag," all have the same s-values, so that summations along such a line give all the frequencies having the particular s-value recorded at the right or the left of the scatter diagram; and the value printed in the lower right corner of each cell is the value of the product $\xi\zeta$ for that cell.

SECTION 6. THE REGRESSION EQUATION AND THE ANALYSIS OF VARIANCE

We have noted [10:16] that having two correlated variables, X_0 and X_1, the total variance V_0 is divisible into two independent parts, one being capable of being estimated from X_1 and the other, $V_{0.1}$, being independent of X_1, thus

$$V_0 = V_{\bar{0}} + V_{0.1} \equiv V_{0\Delta 1} + V_{0.1} \quad . \; . \; . \quad \text{See[10:16]}$$

$$(N-1) = 1 + (N-2) \qquad \text{Degrees of freedom equation} \qquad [10:32]$$

The degrees of freedom of these three variances are $(N-1)$, 1, and $(N-2)$ and, as always, a degrees of freedom equation corresponding to a variance equation can be written. As shown in Chapter VI, the number of degrees of freedom of V_0 is $(N-1)$. Clearly, having a given set of X_1 values the variance of $\overline{X}_0$ (identical with the variance of $\overline{x}_0$) is consequent to one constant, b_{01}, and thus has one degree of freedom. The residual variance, $V_{0.1}$, must then have exactly $(N-2)$ degrees of freedom.

Or, we can determine this by noting that the residual deviations, $X_{0.1}$, are those after two linear restrictions (leading to M_0 and b_{01}) have

been placed upon the X_0 measures. Since M_0 is not known a priori, but found from the sample, this constitutes placing the linear restriction,

$$\Sigma X_0 = N\ M_0 \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ [10{:}33]$$

upon the X_0's. Since b_{01} is not known a priori, but found from the sample, this constitutes placing the added linear restriction,

$$\Sigma X_0 X_1 = N\ M_0 M_1 + N\ V_1 b_{01} \quad . \ . \ . \ [10{:}34]$$

upon the X_0's. Having given, i.e., fixed, values of X_1, the summation $\Sigma X_0 X_1$ is of course linear in X_0. When fitting parabolic regression lines further linear restrictions of the sort, $\Sigma X_0 X_1^2$, $\Sigma X_0 X_1^3$, etc., are involved.

The sample at hand is to be conceived as one of an indefinitely large number, all of which have identically the same set of X_1 values. Thus the parent population varies only in having different X_0 values from sample to sample.

The two parameters whose distributions we are concerned with, should many samples be taken, are M_0 and b_{01}, as indicated in [10:33] and [10:34]. Let the first hypothesis that we desire to test be that M_0 is a chance deviate from $\tilde{M}_0$, and the second that b_{01} is a chance deviate from $\tilde{b}_{01}$. We make the first test under the assumption that $\tilde{b}_{01} = b_{01}$ and the second under the assumption that $\tilde{M}_0 = M_0$. Otherwise expressed, we test the hypothesis that M_0 is a chance deviate from $\tilde{M}_0$ for the universe of samples having the regression b_{01} and we test the hypothesis that b_{01} is a chance deviate from $\tilde{b}_{01}$ for the universe of

samples having the mean M_0. Actual equation [10:35] and theoretical equation [10:36] are the basis for the first test and equations [10:35] and [10:37] are the basis for the second.

$$X_0 - M_0 = b_{01}x_1 + x_{0.1} \quad . \; . \; . \; . \; . \; . \quad [10:35]$$

$$X_0 - \tilde{M}_0 = b_{01}x_1 + \tilde{x}_{0.1} \quad \text{Hypothesis when } \tilde{M}_0 \text{ is being tested} \quad [10:36]$$

$$X_0 - M_0 = \tilde{b}_{01}x_1 + \tilde{x}'_{0.1} \quad \text{Hypothesis when } \tilde{b}_{01} \text{ is being tested} \quad [10:37]$$

Subtracting [10:36] from [10:35] and transposing terms, we obtain

$$\tilde{x}_{0.1} = (M_0 - \tilde{M}_0) + x_{0.1}$$

$\tilde{x}_{0.1}$ is divided into two independent parts and for the first sample we have

$$\tilde{V}_{0.1} = (M_0 - \tilde{M}_0)^2 + V_{0.1}$$

in which $(M_0 - M_0)^2$ has one degree of freedom and $V_{0.1}$ has $(N-2)$, so we obtain the variance ratio

$$F_{1(N-2)} = \frac{(M_0 - \tilde{M}_0)^2 \; (N-2)}{V_0(1-r_{01}^2)} \quad \begin{array}{l} F \text{ to test hypo-} \\ \text{thesis true mean} \\ = \tilde{M}_0 \text{ when true} \\ \text{regression} = b_{01} \end{array} \quad [10:38]$$

Utilizing [10:35] and [10:37] we similarly obtain

$$F_{1(N-2)} = \frac{[(b_{01} - \tilde{b}_{01})^2 \; V_1] \; (N-2)}{V_0(1-r_{01}^2)} \qquad [10:39]$$

F to test hypothesis true regression $= \tilde{b}_{01}$ when true mean $= M_0$

The qualifications next to formula numbers [10:38] and [10:39] "when true regression = b_{01}" and "when true mean = M_0" may ordinarily be considered immaterial because in general the correlation between mean and regression coefficient approximates zero.

The square root of $F_{1(N-2)}$ is "Student's t" and if $r < .90$ and $N > 25$ (in fact, for most purposes if $N > 12$) it may be interpreted as a deviate in a unit normal distribution.

The one-third-sigma rule for the number of decimal places to keep when publishing, given in Chapter VI, Section 7, is based upon the value of the standard error of the item in question. The square root of a variance ratio is a critical ratio, so from [10:38] we note that

$$V_{M_0} = \frac{V_0(1-r_{01}^2)}{N-2} \quad \text{Variance error of } M_0 \text{ determined from correlated data} \quad [10:40]$$

From this we obtain the standard error and apply the one-third-sigma rule. For the Galton data,

$$V_{M_1} = \frac{2.1108(1-.3897^2)}{322} = .00559; \; \sigma_{M_1} = .075;$$

$$\text{and } \frac{\sigma_{M_1}}{3} = .02.$$

We accordingly publish M_1 as 68.31.

From [10:39] we note that

$$V_{b_{01}} = \frac{V_0(1-r_{01}^2)}{V_1(N-2)} \quad \text{Variance error of } b_{01}, \text{ cf [13:143]} \quad [10:41]$$

This, with formulas [6:58] and [6:60] provides the basis for expressing published results as herewith:

$M_1 = 68.31$; $V_1 = 2.11$; $\sigma_1 = 1.45$; $b_{12} = .26$

$M_2 = 68.22$; $V_2 = 4.9$; $\sigma_2 = 2.21$; $b_{21} = .59$;

and $r_{12} = .39$.

SECTION 7. THE TRUSTWORTHINESS OF A CORRELATION COEFFICIENT

The [] term in [10:39] $=[r_{01}\sigma_0 - \text{true}(r_{01}\sigma_0)]^2$. If, instead of assigning a hypothetical value to the product of r_{01} and σ_0, we assign such a value to r_{01} and consider all samples having the standard deviation σ_0, then the right-hand member of [10:39] becomes

$$\frac{(r_{01} - \tilde{r}_{01})^2/(N-2)}{1-r_{01}^2}$$

which, at first sight, looks as though it provided an $F_{1(N-2)}$ test for r_{01}, but this is not so because, when fixing the value of σ_0, we have imposed the condition $\Sigma x_0^2 = NV_0$ and this is not linear in the x_0's.

However, in the special case in which the hypothesis is that $\tilde{r}_{01} = 0$, there are such cancellations that [10:39], or the similar expression, interchanging variables, becomes

$$F_{1(N-2)} = \frac{r_{01}^2(N-2)}{1-r_{01}^2} \qquad \text{F to test hypothesis that true correlation} = 0 \qquad [10:42]$$

and we have a sound test of the significance of $(r_{01}-0)$.

We will mention three unequally excellent procedures which are available when $\tilde{r}_{01}$ does not

equal zero. For the most precise evaluation for samples of 25 or smaller we may use David's *Tables of the ordinates and probability integral of the distribution of r in small samples (1938).*

A second method, which has valuable combinatorial properties in addition to excellent precision in the interpretation of single correlation coefficients, is the r-into-z technique of Fisher (1920-21).* Using "*ln*" to designate a natural logarithm and "log" to designate one to the base 10, the transformation is

$$z = \frac{1}{2}\left[ln(1+r) - ln(1-r)\right] = \frac{1}{2}\, ln\, \frac{1+r}{1-r} \qquad [10:43]$$

Fisher's r-into-z transformation

Since *ln* a = 2.3025851 log a, we may write this using logarithms to the base 10:

$$z = 1.1512925\ [\log(1+r) - \log(1-r)] = 1.1512925\ \log\frac{1+r}{1-r} \qquad [10:43a]$$

Fisher has shown that z has the happy property of being very nearly normally distributed with a standard deviation which is a function of the size of the sample only:

$$\sigma_z = \frac{1}{\sqrt{N-3}} \qquad [10:44]$$

Standard error for all values of z

If we postulate the value $\tilde{r}$, we apply the r-into-z transformation to both r and $\tilde{r}$, obtaining

*Herein is a table of z for values of r, and in Fisher, STATISTICAL METHODS FOR RESEARCH WORKERS (1925 et seq.) is a table of r for values of z. A more detailed table of this relationship is given in column "u" and "tanh u" of Table II of SMITHSONIAN MATH. TABLES, Hyperbolic Functions (1911 et seq.)

z and $\tilde{z}$. The critical ratio that concerns us is

$$\frac{z-\tilde{z}}{\frac{1}{\sqrt{N-3}}} = (z-\tilde{z})\sqrt{N-3} \quad \text{Normally distributed critical ratio to test } (r-\tilde{r}) \qquad [10:45]$$

The third and older method is to obtain a critical ratio by dividing r by its standard error. This standard error is frequently taken as $= (1-r^2)/\sqrt{N}$, but an improved result is gotten if N is replaced by the number of degrees of freedom, $N-2$ (though it is true that one of the two degrees of freedom lost does not represent a linear restriction).

$$\sigma_r = \frac{1-r^2}{\sqrt{N-2}} \quad \text{Standard error of a total correlation coefficient} \qquad [10:46]$$

Many derivations of σ_r have been made and with slightly different outcomes. Perhaps the best first approximation is

$$\sigma_r = \frac{1-r^2}{\sqrt{N-2-4.5r^2}} \quad \cdots\cdots\cdots \qquad [10:46a]$$

which is obtained by combining two of Romanowsky's formulas.*
The critical ratio that now concerns us is

$$\frac{r-\tilde{r}}{\sigma_r} = \frac{(r-\tilde{r})\sqrt{N-2}}{1-r^2} \quad \text{Non-normally distributed critical ratio to test } (r-\tilde{r}) \qquad [10:47]$$

If N is large (say $N > 15$) or r small (say $r < .9$),

* U. Romanowsky, On the moments of standard deviation and of correlation coefficients in samples from a normal population, METRON, Vol.5, no.4, 31-XII, 1925,-formulas [121] and [124]

and particularly if both of these are true, [10:47] will give very serviceable results if we assume r to be normally distributed, but otherwise the lack of knowledge of the form of distribution of [10:47] is serious.

For the Galton data, where variables were labeled X_1 and X_2, we can test the hypothesis that there is zero correlation between mid-parent height by [10:42]

$$F_{1,322} = \frac{(.3897)^2 \; 322}{1 - (.3897)^2} = 57.67$$

This is so large a squared critical ratio that there is negligible probability that it could have arisen as a matter of chance, thus disproving the hypothesis.

Suppose some theory of genetics asserts that the correlation should be 4/9 and we desire to test this hypothesis. A test of [.3897 - .4444] would not take cognizance of the fact that the 4/9 is a value assuming no grouping error, whereas the .3897 has involved standard deviations in which, according to Sheppard's correction for coarseness of grouping [6:49], there is an appreciable error. We could correct the observed correlation coefficient for coarseness of grouping and then compute F, but this would not be precise for the entire theory of F and its distribution involved no such procedure. We must therefore keep the observed value .3897 exactly as it is and alter the theoretical value, 4/9, by an amount which compensates for the coarseness of grouping. Sheppard's correction for the variance is

$${}_cV = V - \frac{i^2}{12} = V\left(1 - \frac{i^2}{12V}\right)$$

Sheppard's correction to V for coarseness of grouping [see 6:49]

in which V is the observed value and i the size

of the grouping interval of the X's. If we take the reducing factor $[1-(i^2)/(12V)]$ as a serviceable estimate of that to apply to the theoretical situation, we have

$$ {}_c\tilde{r}^2_{12} = \frac{\tilde{c}^2_{12}}{{}_c\tilde{V}_1 \; {}_c\tilde{V}_2} = \frac{\tilde{c}^2_{12}}{\tilde{V}_1\left(1-\dfrac{i_1^2}{12V_1}\right)\tilde{V}_2\left(1-\dfrac{i_2^2}{12V_2}\right)} $$

yielding the theoretical r^2 value, stepped down for coarseness of grouping,

$$ \tilde{r}^2_{12} = \frac{\tilde{c}^2_{12}}{\tilde{V}_1 \; \tilde{V}_2} = {}_c\tilde{r}^2_{12}\left(1-\frac{i_1^2}{12V_1}\right)\left(1-\frac{i_2^2}{12V_2}\right) \qquad [10:48] $$

For the Galton data $i_1 = i_2 = 1.0$ (not .5 as was the case in the computation involving ξ and ζ); $V_1 = 2.1108$; $V_2 = 4.8765$; and ${}_c\tilde{r}_{12} = 4/9$. We accordingly obtain $\tilde{r}_{12} = .4318$, and the difference that concerns us is $(r_{12} - \tilde{r}_{12}) = (.3897 - .4318)$. Transforming r and $\tilde{r}$ into z and $\tilde{z}$, we obtain, $z = .4114$, and $z = .4621$, so that the [10:45] critical ratio is $(.4114 - .4621)/\sqrt{321}$, which $= -.9084$. The probability that a positive or negative difference as great as this would arise as a matter of chance if r is a chance deviate from $\tilde{r}$ is .36. Thus the data are in excellent conformity with the hypothesis ${}_c\tilde{r}$ (i.e., $\tilde{r}$ without a grouping error) $= 4/9$.

By [10:47] the critical ratio is $-.8907$ and the corresponding P assuming normality is .37, which though in error is not sufficiently so to lead to a different practical conclusion.

SECTION 8. AVERAGING CORRELATION COEFFICIENTS

To illustrate the combinatorial excellence of z from r let us assume that a number of determinations of the correlation between mid-parent height and offspring height were made and that, except for the numbers of cases involved, they were equally trustworthy. Assume the results to be

an r of .3897 from a sample of 324
an r of .3542 from a sample of 156
an r of .4421 from a sample of 13

An r based upon all the data is desired. The procedure is to transform each r into a z, weight these inversely as their variance errors, and average. Thus

$r = .3897 \rightleftharpoons z = .4114$, which is to be weighted 321
$r = .3542 \rightleftharpoons z = .3702$, which is to be weighted 153
$r = .4421 \rightleftharpoons z = .4748$, which is to be weighted 10

$$\text{Mean } z = \frac{321(.4114) + 153(.3702) + 10(.4748)}{321 + 153 + 10}$$

$$= .3997 \rightleftharpoons r = .3797.$$

This mean z and equivalent r are as reliable as if computed from a sample of 487, which = 321 + 153 + 10 + 3.

SECTION 9. THE TRUSTWORTHINESS OF A POINT ON THE REGRESSION LINE

The regression [10:01] enables us to estimate X_0 knowing X_1. In this case the error of estimate is $(X_0 - \bar{X}_0)$ and the variance error is

$$V_{0.1} = V_0(1 - r_{01}^2)$$

Variance of the errors of estimate when X_0 is estimated from X_1,--see [10:09] and [10:24]

The square root of this is the standard error of estimate. It is the crucial measure giving the precision with which we can estimate one variable from a knowledge of another. Since the quantities $(X_0 - \overline{X}_0)$ are "errors" we may regularly assume them to be normally distributed with a mean of zero and a standard deviation of

$$\sigma_{0.1} = \sqrt{V_{0.1}} = \sigma_0 \sqrt{1-r_{01}^2} \qquad [10:49]$$

Standard error of estimate (See [10:09] and [10:24])

If our concern is not with individual X_0 scores, but with the difference between the regression line found and the true regression line, then we are interested in the variance errors of M_0 and b_{01} as already discussed. We may also be interested in the trustworthiness of some specific $\overline{X}_0$ corresponding to a specific X_1. The variance error of this has been given by Pearson (1913 Freq.) and, with minor modification of changing N to $N-2$, is

$$V(\overline{X}_0) = \frac{V_0(1-r_{01}^2)}{N-2}\left[1+\frac{(X_1-M_1)^2}{V_1}\right] \qquad [10:50]$$

Variance error of a point on a regression line when distributions of arrays are similar

We can illustrate the use of this formula by citing a problem arising in the "standardization" of educational or psychological tests. Let X_1 be the score on a comprehensive and highly reliable* psychological or educational test which is applicable to a wide range of talent, which we will call the "anchor test" and for which adequate

* If not highly reliable, a modification of the procedure here outlined, incorporating the reliability of the X_1 scores is necessary.

"norms" exist. Let X_0 be a score upon a new test in the process of being "standardized," or having norms established. The correlation between it and X_1 for a sample of N has been obtained. We desire to tie the new test to the anchor test and secure a set of parallel scores such that we can assert $X_{0\Delta 1a}$ is the norm for those of ability X_{1a}; $X_{0\Delta 1b}$ for those of ability X_{1b}; etc. Introducing this value X_{1a} into [10:01] and noting that $X_{0\Delta 1} \equiv \overline{X}_0$, yields $X_{0\Delta 1a}$ which is the desired parallel score, or equivalent score, to X_{1a}. Substituting X_{1a} for X_1 in [10:50] gives the variance error of this equivalent score.

Clearly, if r_{01} is high and X_{1a} does not differ greatly from M_1, we obtain $X_{0\Delta 1a}$ with high precision even though N is rather small. This is thus a means, at moderate testing cost, of establishing the norms of a new test by tying them to those of highly reliable previous measure.

SECTION 10. CORRELATION BETWEEN RANKS

A simple formula for the correlation between ranks is readily derived from the product-moment formula by expressing the covariance as a function of the variance of the differences of paired measures x_1, x_2.

$$V(x_1 - x_2) = V_1 + V_2 - 2c_{12} = V_1 + V_2 - 2\sigma_1\sigma_2 r_{12} \qquad [10:51]$$

Thus

$$r_{12} = \frac{V_1 + V_2 - V(x_1 - x_2)}{2\,\sigma_1\,\sigma_2} \qquad [10:52]$$

Difference formula for product moment r

We record, as being at times convenient, a similarly derived formula based upon the variance of the sum of paired measures;

$$r_{12} = \frac{V(x_1 + x_2) - V_1 - V_2}{2\,\sigma_1\,\sigma_2}$$ Sum formula for product moment r [10:53]

or combining these two

$$r_{12} = \frac{V(x_1 + x_2) - V(x_1 - x_2)}{4\,\sigma_1\,\sigma_2}$$ Sum and difference formula for product moment r [10:54]

If, as in the case of ranked data involving no missing ranks from 1 to N, we have $V_1 = V_2$, formula [10:52] simplifies. From [14:128] giving the sum of N numbers, 1, 2, 3, . . . , N, and of their squares, we have, if X stands for the successive ranks 1, 2, 3 . . . , N,

$$\Sigma X = \frac{N^2}{2} + \frac{N}{2} = \frac{N(N+1)}{2}$$

so that

$$M_x = \frac{N+1}{2}$$ The mean of N ranks [10:55]

Also

$$\Sigma X^2 = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6} = \frac{N(2N+1)(N+1)}{6}$$

so that

$$\text{Mean } X^2 = \frac{(N+1)(2N+1)}{6}$$ [10:56]

$$V_x = (\text{mean } X^2 - M_x^2) = \frac{N^2-1}{12}$$ The variance of N ranks [10:57]

Let d stand for the differences in paired ranks, i.e., for $(x_1 - x_2)$ of formula [10:52] when the scores in question are ranks, and let ρ_{12} be the

product moment correlation between ranks, then, utilizing [10:52] and [10:57]

$$\rho_{12} = 1 - \frac{6\ \Sigma d^2}{N(N^2 - 1)} \qquad \text{Spearman rank correlation coefficient} \qquad [10:58]$$

This formula must not be confused with Spearman's "foot-rule" rank correlation formula, which, though simpler, has such shortcomings as never to be preferred to [10:58].

The coefficient ρ_{12} is exactly a product moment correlation coefficient, when scores are ranks, and if used in a regression equation in which the predictor variable is expressed as a rank, it should not be altered in any way. In a theoretical treatment, wherein the assumption is made that the real distributions underlying the ranks are normal, Pearson (1907) gives a formula for estimating from ρ the r that would have been obtained had the continuous normally distributed variables been used. It is

$$r = 2 \sin\frac{\pi}{6}\rho \qquad \text{Pearson's correction to Spearman's } \rho \qquad [10:59]$$

This correction is small and in general it suffices to report ρ and its standard error. The standard error of ρ and of r-from-ρ have been given by Pearson and are, after first substituting $(N-2)$ for N, to allow for the loss of two degrees of freedom, as herewith

$$\sigma_\rho = \frac{1-\rho^2}{\sqrt{N-2}}(1 + .086\ \rho^2 + .013\ \rho^4 + .002\ \rho^6) \qquad [10:60]$$

Standard error of Spearman's ρ

$$\sigma_r = 1.0472\frac{1-r^2}{\sqrt{N-2}}(1 + .042r^2 + .008\ r^4 + .002\ r^6)$$

Standard error of r from ρ [10:61]

If one of the paired measures is a normally

distributed variate and the other a series of ranks, though the underlying variable is in fact normally distributed, and if the product moment correlation between them is called ρ', Pearson (1914) gives the following formula to estimate the correlation had the normally distributed scores been available:

$$r = \sqrt{\frac{\pi}{3}}\,\rho' = 1.0233\,\rho' \quad \ldots\ldots\ldots \quad [10:62]$$

Correction in ρ for ties in ranks. It frequently occurs that two or more ranks are tied. In this case all of the tied ranks are given a single value, which is the average of the ranks represented by the tied measures. For example, the 11 scores

10, 9, 8, 8, 7, 6, 6, 6, 5, 4, 4

are given rank values

1, 2, 3.5, 3.5, 5, 7, 7, 7, 9, 10.5, 10,5

The mean of these is exactly the value given by [10:55], but mean x^2 for these is not identical with the answer given by [10:56], and V_x is not identical with the value given by [10:57]. Horn (1942) has noted that the correction in [10:57] is simple and has given tables and approximations to facilitate computation when there are different numbers of ranks in the ties.

We note that if a number of ranks X, $X+1$, $X+2$, . . . , $X+k$ are replaced by their average, $X+.5k$, and if their squares X^2, $(X+1)^2$, . . . $(X+k)^2$ are replaced by $(X+.5k)^2$, $(X+.5k)^2$, . . . , $(X+.5k)^2$, the sum of the latter squares is less than the former by an amount which is equal to k times the variance of k ranks, thus

$$\sum_{X=X}^{X=X+k} X^2 = \sum^{k}(X+.5k)^2 + \frac{k(k^2-1)}{12} = k(X+.5k)^2 + \frac{k(k^2-1)}{12} \quad \ldots\ldots\ldots [10:63]$$

We may rewrite [10:58] thus

$$\rho_{12} = 1 - \frac{\Sigma d^2}{2\sqrt{\dfrac{N(N^2-1)}{12}}\sqrt{\dfrac{N(N^2-1)}{12}}} = 1 - \frac{\Sigma d^2}{2\sqrt{NV_1}\ \sqrt{NV_2}}$$

Herein the V's are given by [10:57], but if there are tied ranks, NV_1 is too large by an amount equal to the sum of all the functions $k(k^2-1)/12$ for all the ties in the first series and similarly NV_2 is too large by the sum of all the functions $[k(k^2-1)/12]$ for the second series. To illustrate in the case of the 11 ranks just given: the sum of the squares of the measures 1, 2, 3, 5, 3.5, 5, 7, 7, 7, 9, 10.5, 10.5 is equal to the sum of the squares of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 minus $\{[2(2^2 - 1)/12] + [3(3^2 - 1)/12] + [2(2^2-1)/12]\}$, because of ties 3.5, 3.5 and 7, 7, 7 and 10.5, 10.5. The deduction is 3(=.5+2+.5). We accordingly, in this instance, deduct 3 and set

$$NV_1 = \frac{11(11^2-1)}{12} - 3 = 107$$

Of course, a similar correction would be necessary for NV_2 so that the formula for ρ becomes

$$\rho_{12} = 1 - \frac{6\ \Sigma\ d^2}{\sqrt{N(N^2-1)-Sk_1(k_1^2-1)}\ \sqrt{N(N^2-1)-Sk_2(k_2^2-1)}}$$

ρ corrected for ties in ranks [10:64]

Herewith is a short table of $k(k^2-1)$ values for different numbers of ties in ranks:

TABLE X C

CORRECTIONS WHEN COMPUTING ρ FROM TIED RANKS

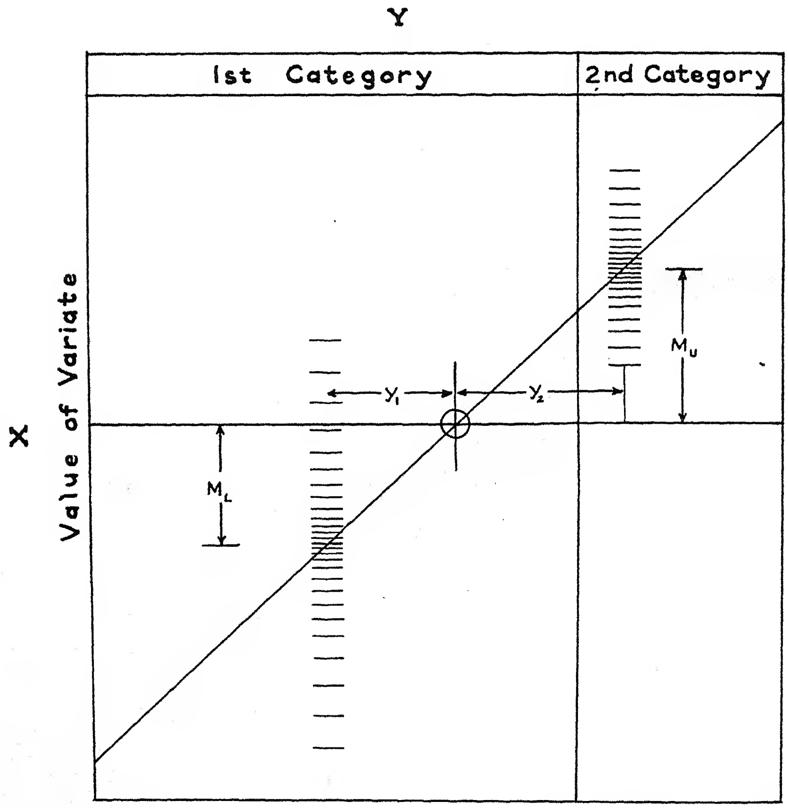
k: No. of ranks tied	2	3	4	5	6	7	8	9	10
$k(k^2-1)$	6	24	60	120	210	336	504	720	990

SECTION 11. BISERIAL CORRELATION

A frequent correlation situation is one in which one of the paired variables is dichotomous. Chart XII illustrates this case. X is the gradu-

CHART X II

A BISERIAL SCATTER DIAGRAM



ated variable and Y the two-category variable. With such data, in addition to the straightforward problems of estimating X knowing Y and of estimating Y knowing X, we may be interested in estimating the correlation and the regressions maintaining between X and Y', a continuous normally distributed variable which has been forced into the dichotomy given by Y.

The essential statistics obtained from a scatter diagram are shown in Chart X II.

The units of measurement of X may well be those of the raw scores, grouped in not less than 12 classes if desired. In the case of Y, simplicity results if the lower score is called 0 and the higher score called 1. If the proportion of cases receiving a 0 score is q, then the constants of the Y distribution are:

$$M_y = p \quad . \; . \; . \; . \; . \; . \; . \; . \quad [10:65]$$

$$V_y = pq \quad . \; . \; . \; . \; . \; . \; . \; . \quad [10:66]$$

For the X series we compute

$$M_x \; ; \; V_x$$

The number of cases is N and the covariance is

$$c_{xy} = \frac{\Sigma XY}{N} - M_x M_y = \frac{\sum^{Nq} X_\ell \times 0 + \sum^{Np} X_u \times 1}{N} - M_x p$$

$$= pq(M_u - M_\ell) \quad . \; . \; . \; . \; . \; . \; . \quad [10:67]$$

in which the X_ℓ measures are those paired with a zero Y score and the X_u measures are those paired with a Y score of 1. Also M_ℓ is the mean of the X measures in the lower group and M_u the mean of the X measures in the upper group.

$$b_{xy} = M_u - M_\ell \quad \text{Regression of continuous variable upon two-category variable} \quad [10:68]$$

$$b_{yx} = \frac{pq(M_u - M_\ell)}{V_x} \quad \text{Regression of two-category variable upon continuous variable} \quad [10:69]$$

$$r_{xy} = \frac{(M_u - M_\ell)\sqrt{pq}}{\sigma_x} \quad \text{Biserial product moment } r \quad [10:70]$$

$$\bar{X} = (M_u - M_\ell)\,Y + M_\ell \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad [10:71]$$

$$\bar{Y} = \frac{(M_u - M_\ell)\,pq}{V_x} X + p - \frac{(M_u - M_\ell)\,pq}{V_x} M_x \quad [10:72]$$

The variance ratio test of the regression coefficient $(M_u - M_\ell)$ is given by [10:39], and the variance ratio test of M_x by [10:38].

If N is small, these tests, [10:39] and [10:38], for the regression $(M_u - M_\ell)pq/V_x$ and the mean, M_y $(=p)$, should be applied with misgiving because the variances involved derive from two-point distributions.

A test of $(M_u - M_\ell)$ alternative to [10:39] is the critical ratio test $(M_u - M_\ell)/\sigma_{M_u - M_\ell}$, which squared is,

$$F_{1,\,N-2} = \frac{(M_u - M_\ell)^2}{\dfrac{V_u}{N_u - 1} + \dfrac{V_\ell}{N_\ell - 1}} \qquad \text{(See [6:19])} \quad [10:73]$$

If, in the population $\tilde{V}_u = \tilde{V}_\ell$, then approximately

$$\frac{V_u N_u}{N_u - 1} = \frac{V_\ell N_\ell}{N_\ell - 1}$$

and the just recorded squared critical ratio becomes

$$F_{1,N-2} = \frac{(M_u - M_\ell)^2 \; pq(N-2)}{V_x(1-r^2_{xy})}$$

which is exactly the [10:39] variance ratio test of the regression coefficient $(M_u - M_\ell)$. Thus, if the data are homoscedastic (*equal variability of arrays under the definition of equal variability of*

$$V_u\left(\frac{N_u}{N_u-1}\right) = V_\ell\left(\frac{N_\ell}{N_\ell-1}\right)$$

and not the older definition, $V_u = V_\ell$), the [10:39] test is adequate, but if V_u does not approximately equal V_ℓ the squared critical ratio test [10:73] is the more precise.

Estimates of biserial relationships, assuming the dichotomy is one forced upon an underlying normal variate. Under this assumption we can compute the correlation and the regression equations, but as the assumption itself is always open to serious questioning, though how serious we cannot express quantitatively or in probability terms, a computation of the standard errors of these modified correlation and regression constants is objectionable. It could only lead to fiducial results in which the presumably major source of error (assumption of normality) was neglected while the minor source (small size of sample) was treated as though it were the sole source of error.

If the Y measures recorded in the two-category manner had derived from a normal distribution having a mean of zero and a variance of one, the means of the lower and upper categories would be

$-z/q$ and z/p respectively, in which z is the ordinate in a unit normal distribution at that point for which the proportion of cases to the left is q, and the proportion to the right is p, as given by [8:25] and [8:26]. Further, the regression of X upon this normal variate, which we designate y' (the same as x of tables of the unit normal distribution), will be a line passing through the intersection of M_ℓ and $-z/q$ and the intersection of M_u and z/p. In Chart X II, the sum of the absolute values of the distances $\bar{x}_2$ and $\bar{x}_1$ equals $(M_u - M_\ell)$, and the sum of the absolute values of the distances y_2 and y_1 is, in terms of the unit normal distribution,

$$\frac{z}{p} + \frac{z}{q} \text{ which } = \frac{z}{pq}, \text{ so that}$$

$$b_{xy'} = \frac{(M_u - M_\ell)\, pq}{z} \qquad \text{Regression of actual continuous variate upon assumed normal variate} \qquad [10:74]$$

Of course $V_{y'} = 1$, so that

$$r_{xy'} = \frac{(M_u - M_\ell)\, pq}{\sigma_x z} \qquad \text{Biserial } r \qquad [10:75]$$

The designation of this as "biserial r" follows long established practice, but a more accurate designation would be "biserial r corrected for an enforced dichotomy upon a normal variate."

The regression equation of X upon y' serves no practical need since one does not actually have y'. The other regression is

$$\bar{y}' = \frac{(M_u - M_\ell)\, pq}{V_x\, z}(X - M_x) \quad \ldots\ldots\ldots\ldots \quad [10:76]$$

The variance error of estimate of y' *(assuming zero error in the assumption of normality of y')* is

$$V_{y'.x} = 1 - r^2_{xy'}, \quad \ldots\ldots\ldots \quad [10:77]$$

The variance error of biserial r, as given by Soper (1914), is

$$V_{\text{biser.}r} = \frac{1}{N}\{\frac{pq}{z^2} - [\frac{3}{2} + (1 - \frac{py'}{z})(1 + \frac{qy'}{z})]$$

$$(\text{biser.}r)^2 + (\text{biser.}r)^4\} \quad \cdot\cdot \quad [10:78]$$

but of course this variance error does not have incorporated in it the error because of the assumption of normality. For dichotomies wherein q is not less than .05, a close approximation to [10:78] is

$$V_{\text{biser.}r} = \frac{[\frac{\sqrt{pq}}{z} - (\text{biser.}r)^2]^2}{N} \quad \ldots \quad [10:79]$$

A common practice in the statistical analysis of true-false, or right-wrong, test items has been to judge of their excellence by the size of the biserial correlations between them and a graduated criterion, or by the critical ratios biser.$r/\sigma_{\text{biser.}r}$. In view of the fact that the continuous variable y' is not actually available, nor presumably ever will be available for it is not contemplated that the future scoring of such an item will ever be other than "right" or "wrong", and in view of the shortcomings of the formula for the standard error of biserial r,

it would seem better to employ biserial product moment r and the precise tests connected therewith. The hazards of employing biserial r instead of a real product moment r, such as is biserial product moment r, are increased if several items of a test are combined in a multiple regression equation to estimate a graduated criterion, (a) because the best weights, in the minimal variance error of estimate sense, derive from the product moment r's and cannot be improved upon in any manner, except by resort to higher order, or curvilinear regression, and (b) because the precise tests existing for product moment multiple and partial correlation do not apply if correlation coefficients other than product moment coefficients are employed. The observations in Section 12,—tetrachoric r,—upon dichotomies and range of talent apply in biserial r situations.

The writer has been kindly supplied with the data of Table X D, which arose in the experimental work of L. R. Waldron, State College Station, Fargo, North Dakota.

TABLE X D

CERTAIN BISERIAL DATA

X	Y=0 No. of cases	Y=1 No. of cases	Resulting constants
10		2	$M_u = 7.55$
9		4	$M_\ell = 3.05$
8		15	$\sigma_x = 2.367$
7		12	$p = \frac{2}{3}$
6		8	
5	3	1	
4	4		
3	7		$z = .3636$
2	5		Product moment biserial $r = .90$
1	2		
	21	42	Biserial $r = 1.13$

Here the assumption of normality in the Y variate is so untenable as to lead to an absurd biserial r. Should a situation be only a little less extreme than this, it would not be revealed by yielding an impossible r value. The student should make the assumption of normality only after careful consideration and he should not expect an unsoundness in this respect to be automatically revealed by absurd statistical outcomes.

It seems to the writer fairly satisfactory to use biserial r as a terminal statistic in connection with the data of Table X E drawn from psychological examination in the United States Army (see Yerkes 1921).

TABLE X E

SCHOOL ATTAINMENT AND ARMY ALPHA SCORES

SCORE IN ARMY ALPHA INTELLIGENCE TEST	NUMBER OF MEN WHO LEFT SCHOOL BELOW THE 9TH GRADE	NUMBER OF MEN WHO LEFT SCHOOL ABOVE THE 8TH GRADE
205-212		1
200-204		3
195-199		14
190-194		17
185-189	1	49
180-184	2	54
175-179	8	78
170-174	12	126
165-169	18	149
160-164	15	200
155-159	20	244
150-154	45	305
145-149	58	352
140-144	74	338
135-139	101	407
130-134	145	507
125-129	190	528
120-124	216	530
115-119	317	643
110-114	393	674

TABLE X E
(CONTINUED)
SCHOOL ATTAINMENT AND ARMY ALPHA SCORES

SCORE IN ARMY ALPHA INTELLIGENCE TEST	NUMBER OF MEN WHO LEFT SCHOOL BELOW THE 9TH GRADE	NUMBER OF MEN WHO LEFT SCHOOL ABOVE THE 8TH GRADE
105-109	507	682
100-104	582	691
95- 99	761	712
90- 94	908	725
85- 89	993	769
80- 84	1,181	693
75- 79	1,371	642
70- 74	1,604	648
65- 69	1,709	567
60- 64	1,962	581
55- 59	2,249	430
50- 54	2,272	346
45- 49	2,429	305
40- 44	2,455	229
35- 39	2,473	200
30- 34	2,490	154
25- 29	2,213	106
20- 24	1,835	60
15- 19	1,511	42
10- 14	545	13
5- 9	432	5
0- 4	183	3
	34,280	13,822
	13,822	
	48,102	

$M_u = 98.758$ $M_l = 54.987$

$\sigma_x = 36.606$ $p = .28735$

$z = .34083$

biserial product moment $r = .541$

biserial $r = .718$

Considerations in support of the utility of this biserial correlation are *(a)* the naive concept of the correlation between intelligence test score and schooling would assuredly involve the amount of schooling as a graduated variable, *(b)* the future actual treatment of schooling as a graduated variable is entirely possible, *(c)* the distribution of "highest school grade attended" is probably not radically non-normal, *(d)* the correlation is so high and the sample so large that a precise test of the significance of $(r_{biser.}-0)$ is not needed, and *(e)* presumably this statistic is a terminal statistic and will not be incorporated into further algebraic treatment.

SECTION 12. TWO-BY-TWO-FOLD CORRELATION

Phi: The issue of continuity which arose in connection with the Y variable in the case of biserial correlation, here arises in connection with both variables. We first ascertain the pertinent statistical constants for the actual point distributions as given by the data. In this case the product moment correlation has been known as Yule's ϕ, and is here designated ϕ. It is, in every detail, a rigorous product moment correlation coefficient, and the use of ϕ instead of r to designate it merely calls attention to the fact that it is computed from a two-by-two-fold. Its computation is simple and its use calls only for a certain reservation in the case of variance ratio or chi-square tests applied to very small samples.

It has been customary to plot two-by-two-fold scatter diagrams as in Charts X III and X IV, so that the proportions p and p' are each equal to or greater than .5, even though this involves divergence from the usual scaling of abscissa from left to right and of ordinate from bottom

to top. The X and Y scales are indicated in Chart X IV.

CHART X III

TWO-BY-TWO-FOLD SCATTER DIAGRAM

				X
	a	b	$a+b$	1
	c	d	$c+d$	0
	$a+c$	$b+d$	N	
Y	1	0		

CHART X IV

TWO-BY-TWO-FOLD SCATTER DIAGRAM
BASED UPON PROPORTIONS

				X
	α	β	p	1
	γ	δ	q	0
	p'	q'	1	
Y	1	0		

$\alpha = a/N$; $\beta = b/N$; $\gamma = c/N$; $\delta = d/N$; $p = (a+b)/N$; $p' = (a+c)/N$. It is readily established that $M_x = p$; $V_x = pq$; $M_y = p'$; $V_y = p'q'$; and the co-variance

$$c_{xy} = \frac{ad-bc}{N^2} = \alpha\delta-\beta\gamma \qquad [10:80]$$

$$\phi = \frac{c_{xy}}{\sigma_x \sigma_y} = \frac{ad-bc}{N^2 \sigma_x \sigma_y} = \frac{\alpha\delta-\beta\gamma}{\sqrt{pq\ p'q'}} \qquad \text{Product moment } r \text{ in a two-by-two-fold} \qquad [10:81]$$

$$b_{xy} = \frac{\alpha\delta-\beta\gamma}{p'q'} \qquad \text{And similarly for } b_{yx} \qquad [10:82]$$

$$\bar{X} = b_{xy}\ Y + p - b_{xy}p' \qquad \text{And similarly for } \bar{Y} \qquad [10:83]$$

$$F_{1,N-2} = \frac{(p-\tilde{p})^2\ (N-2)}{pq\ (1-\phi^2)} \qquad \text{Variance ratio to test } (p-\tilde{p}) \qquad [10:84]$$

This may be compared with the critical ratio test for the mean, available when no second variable Y is utilized. The error variance,— $pq(1-\phi^2)$, may be looked upon as the sum of $(N-2)$ independent variables arising from two-point distributions. The tabled distributions of F, or of Student's t^2, assume that these independent variances arise from normal distributions. For the usual situations in which N is, say, greater than 15, highly serviceable answers result from employing the usual P from F.

$$F_{1,N-2} = \frac{(b_{xy}-\tilde{b}_{xy})^2\ p'q'(N-2)}{pq\ (1-\phi)^2} \qquad \text{Variance ratio to test } (b_{xy}-\tilde{b}_{xy}) \qquad [10:85]$$

The $\tilde{p}$ in [10:84] and the $\tilde{b}_{xy}$ in [10:85] are a priori values, or points of reference in which interested.

For samples of small size precise tests of significance, as illustrated in Chapter IX, Section 2, are preferable to tests based upon [10:84]

and [10:85].

The variance error of ϕ, as derived by Yule (see Pearson and Heron, 1913) is

$$V_\phi = \frac{1}{N}\{1-\phi^2 + (\phi + \frac{\phi^3}{2})(\sqrt{\frac{q}{p}} - \sqrt{\frac{p}{q}})(\sqrt{\frac{q'}{p'}} - \sqrt{\frac{p'}{q'}}) - \frac{3\phi^2}{4}$$

$$[(\frac{q}{p}+\frac{p}{q}-2)+(\frac{q'}{p'}+\frac{p'}{q'}-2)]\} \qquad [10:86]$$

$(q<p$ and $q'<p')$

Variance error of ϕ from a two-by-two-fold table

Tetrachoric correlation: This correlation coefficient is an estimate of the correlation to be expected in a parent normal bivariate distribution which, when forced into a two-by-two-fold, yielded the observed distribution. If an approximately normal bivariate parent may reasonably be assumed to exist, if the issue is theoretical or one concerning this parent and not a predictive problem which must use the available two-category measures, if a precise test of significance is not necessary because the sample is large and the magnitude of the correlation rather than its existence is the matter of importance, and if the correlation coefficient obtained is not to be used in further connections wherein a test of significance will be needed, then tetrachoric r will be a useful statistic. The writer has noted the frequent use of tetrachoric r in multiple regression equations in psychological research, but as these almost always lead to issues requiring tests of significance this practice has serious shortcomings.

Pearson (1900) derived equation [10:87] in which r_t is the tetrachoric correlation coefficient, q, p', and δ are as indicated in Chart X IV, x and z are deviate and ordinate in a unit normal distribution at the point of dichotomy

of the X variable and x' and z' the deviate and ordinate in a unit normal distribution at the point of dichotomy of the Y variable.

$$\frac{\delta - qq'}{zz'} = r_t + xx'\frac{r_t^2}{2!} + (x^2-1)(x'^2-1)\frac{r_t^3}{3!}$$

$$+ (x^3-3x)(x'^3-3x')\frac{r_t^4}{4!}$$

$$+ (x^4-6x^2+3)(x'^4-6x'^2+3)\frac{r_t^5}{5!}$$

$$+ (x^5-10x^3+15x)(x'^5-10x'^3+15x')\frac{r_t^6}{6!}$$

$$+ (x^6-15x^4+45x^2-15)(x'^6-15x'^4+45x'^2-15)\frac{r_t^7}{7!} + \cdots$$

Equation giving r_t, the tetrachoric coefficient of correlation [10:87]

To express the law governing successive coefficients of powers of r_t, let $v_n w_n/n$ be the coefficient of r_t^n, v_n be a function of x, and w_n a function of x'; then v_n may be expressed in terms of v's of a lower order:

$$v_n = xv_{n-1} - (n-1)v_{n-2}$$

and similarly

$$w_n = x'w_{n-1} - (n-1)w_{n-2}$$

$$v_0 = 1,\ v_1 = x$$

and similarly

$$w_0 = 1,\ w_1 = x'$$

Thus the equation as written to the r_t^7 term may be continued to any number of additional terms desired should it not converge rapidly enough to make terms above the r_t^7 negligible.

Extensive tables to facilitate the computation of r_t have been computed by P. F. Everitt, Alice Lee, Margaret Moul, Ethel M. Elderton, A. E. R. Church, E. C. Fieller, and J. Pretorius. These are given in Pearson's *Tables* (1914).

The most extensive and useful of these tables give the proportionate area for different dichotomies in both X and Y maintaining for correlations −.95, −.90, −.85, etc., to .95, 1.00. The interval in x and x' (designated h and k in the tables) is $.1\sigma$. Thus, having any x, x', and δ, one can find r_t by interpolation between two successive tables. The tables have many other uses, including the determining of the proportionate frequency in a normal bivariate distribution in any desired rectangular interval (the sides of the rectangle being parallel to the axes) for any of the correlations −.95, −.90, ..., 1.00.

Extensive diagrams for the computation of r_t have been published by Chesire, Saffir, and Thurstone (1933).

If an assumption of normality for one only of the variables seems reasonable,—the other variable being a genuine dichotomy,—the problem reduces to that of biserial r, the graduated variable now having but two values. These require no correction for grouping for it is not postulated that any continuum underlies this variable. We will use the data of Table X F, also

drawn from Psychological Examining in the United States Army, to illustrate ϕ, biserial r from a two-by-two-fold, and r_t.

TABLE X F

ARMY ALPHA SCORES OF LIEUTENANTS IN MEDICAL AND NON-MEDICAL DEPARTMENTS

		Score on army intelligence alpha test: A or B	Score on army intelligence alpha test: Below B	
First Lieutenants	Departments other than Medical	2940	431	3371
First Lieutenants	Medical Department	1799	590	2389
		4739	1021	5760

TABLE X G

FREQUENCIES OF TABLE X F EXPRESSED AS PROPORTIONS

$.5104 = \alpha$	$.0748 = \beta$	$.5852 = p$; $.215215 = x$; $.389809 = z$
$.3123 = \gamma$	$.1205 = \delta$	$.4148 = q$

$.8277 = p'$ $.1773 = q'$

$.925704 = x'$

$.259915 = z'$

Substituting the values given in Table X G for δ, x, x', z, and z' in equation [10:87] and solving for r_t by the method of successive approximations given in Chapter XIV, Section 7, yields $r_t = .277$. Computation of biserial r, assuming a normal distribution to underlie the intelligence test variable, yields $r_{x'y} = .195$. Computation of the product moment biserial correlation yields $\phi = .154$.

To predict the department, knowing the dichotomous army alpha score, calls for the use of ϕ.

The correlation is positive when non-medical departments are at the high end of the scale and the medical department at the low end. To answer the theoretical issue as to the indicated correlation if adequate X and Y variables were available, neither ϕ nor $r_{x'y}$ is very satisfactory. $r_{x'y}$ accepts the point distribution of the "Medical department-other departments" variable as reasonable, and this is certainly very questionable. ϕ accepts the point nature of this variable and also of the intelligence variable, so that it is doubly open to question. The writer is prone to put more confidence in r_t as descriptive of the underlying relationship than in the other two measures, but cannot defend this statistically and would hesitate to attach a standard error to the obtained value .277. (The σ_{r_t} given by [10:88] is .016.

If the assumption of normality for both underlying distributions is unquestionable, the following formula, due to Pearson (1913 Prob.), gives an approximate answer for the variance error of tetrachoric r:

$$V(r_t) = \frac{pqp'q'}{Nzz'}(1-r_t^2)\left[1-\left(\frac{\sin^{-1} r_t}{90^\circ}\right)^2\right] \qquad [10:88]$$

As we may, in general, believe that [10:88] understates the variance error, the result given by it may be taken as a lower estimate and when so taken we find that tetrachoric r is very unreliable when dichotomies are extreme.

If the two-by-two-fold is such that $p=q=p'=q'=.5$, then [10:87] simplifies to

$$r_t = \cos(2\pi\beta) \quad \text{Tetrachoric correlation when dichotomic lines are the medians} \qquad [10:89]$$

This is a useful formula to use in the preliminary investigation of paired graduated measures. Calling cases above the median +, for each vari-

able, and cases below the median −, then a simple count gives the number of pairings of unlike sign. This number divided by N equals 2β, which is U, the proportion of unlike sign pairs, of formula [10:89a].

$$r_t = \cos(\pi U) \qquad \text{Unlike signs formula for } r_t \qquad [10:89a]$$

We may use for the variance error in this case

$$V_{r_t} = \frac{1-r_t^2}{N} 2\pi\,\alpha\beta \qquad \text{Variance error of } r_t \text{ when dichotomic lines are at the medians} \qquad [10:90]$$

In standardized test construction it is common: *(a)* to use a preliminary form of a test having more items (scored right or wrong) than will be retained in the final form; *(b)* to give this preliminary form to an experimental group for the members of which there can be gotten a dichotomous criterion score (if the criterion score is graduated, it is frequently dichotomized at a point at or near the median); and *(c)* to appraise the merits of the separate items by their correlations with the criterion. Yule's ϕ being a product moment correlation is, in the least squared error sense, the best measure of excellence and any other measure, such as is r_t, is a less accurate statement of the relative excellence of the item in its power to predict the dichotomous criterion score. However, r_t has been widely used in this situation and though it has been, in the writer's opinion, generally misused there nevertheless is a real argument in its favor when the experimental group is more homogeneous than the future population to which it is anticipated the test will be given. For example, consider an Air Corps flying aptitude test standardized upon the basis of Air Corps Cadets, but used later upon the more heterogeneous group of Air Corps cadet candidates. An

easy item upon which, say, 90 per cent of cadets and 50 per cent of candidates pass, yields in the experimental group a considerably higher r_t than ϕ. The ϕ reveals the excellence of the item for such samples as the experimental group, whereas r_t will in general be a better estimate than ϕ of the excellence of the item for candidate groups. The general policy of using an item or a test for a different range of talent than that for which there is validation evidence can be cogently criticized, but such practice is at times unavoidable and logical considerations may suggest that r_t, which is an estimate of a situation other than that observed, may have practical merit, in spite of many shortcomings, including the lack of any precise tests of significance. This latter lack is much more serious in factor analysis studies than in item analyses, so the use of r_t's in factor analysis seems to the writer to be inexpedient.

SECTION 13. ADJUSTMENTS FOR COARSENESS OF GROUPING

The problems that arise generally concern theoretical issues of correlation, but practical problems are occasionally present.

We will call the measure being predicted the criterion, or the predictand, and the other measure the predictor or independent variable. In connection with the following regression equation, $\overline{Y}$ is the prediction, Y the predictand, and X the predictor:

$$\overline{Y} = b_{yx} X + M_y - b_{yx} M_x$$

Obviously if X is now, and will continue in the future to be, available in a small number of classes only, there is no point in estimating the correlation, or the regression, that would main-

tain were X a finely graduated measure. In this case no correction for grouping, as regards X, is serviceable.

A similar argument does not always apply to the predictand. Suppose the criterion to be a two-category measure, "superior-inferior," of medical efficiency. Though only two classes are available in the experimental study determining the relationship, one would ordinarily be more concerned with predicting degrees of efficiency than with predicting a two-category status. Thus for the Y variable we may make the most reasonable correction for grouping that is possible. This correction, in the case of a presumably normal distribution underlying a dichotomous criterion, is given by biserial r and the attendant regression equation. Even so, the procedure has the serious drawback that precise tests of significance and precise knowledge of the standard error of estimate are not available. However, if Y' is the continuous variable underlying the dichotomous Y measures, the prediction equation is given by [10:76] and the best, though still questionable, estimate of the variance error of prediction is given by [10:78].

If the number of classes in the criterion is greater than 11, we seldom need to be concerned with a correction for grouping.

If the number of classes in the criterion lies between 5 and 12, we have available one of two corrections depending upon whether the variable is of class indexes or of class means.

Corrections when class indexes are the variable: We will illustrate this problem by the scatter diagrams of Table X H and X I. The proportionate frequencies listed in the cells are those of a normal bivariate distribution with the coarse groupings given. In each instance the correlation for the non-grouped data would be .80, and the variance would be 1.

TABLE X H

NORMAL BIVARIATE DISTRIBUTION IN WHICH THE CORRELATION IS .80, IN CASE OF COARSE GROUPING

			.000017	.000719	.003306	.002168	
		.000049	.003646	.027520	.026076	.003306	
	.000049	.006271	.079945	.127226	.027520	.000719	
.000017	.003646	.079945	.215708	.079945	.003646	.000017	
.000719	.027520	.127226	.079945	.006271	.000049		
.003306	.026076	.027520	.003646	.000049			
.002168	.003306	.000719	.000017				
−3.	−2.	−1.	.0	1.	2.	3.	CLASS INDEXES
−2.8226	−1.8481	− .9206	.0	.9206	1.8481	2.8226	CLASS MEANS

TABLE X I

NORMAL BIVARIATE DISTRIBUTION IN WHICH THE CORRELATION IS .80, IN CASE OF VERY COARSE GROUPING

.000056	.060963	.097637	
.060963	.560760	.060963	
.097637	.060963	.000056	
−2.	.0	2.	CLASS INDEXES
−1.5251	.0	1.5251	CLASS MEANS

First consider the case in which the available variable is the class index. We will designate the variables given by the class indexes, $y_{\rm I}$ and $x_{\rm I}$, and when a statistic is corrected for coarseness of grouping we so indicate by the preceding subscript c, the unavailable continuous variables by $\tilde{y}$ and $\tilde{x}$, true statistics, (that is, of these unavailable variables) by a tilde, and when a statistic is corrected for coarseness of grouping we so indicate by the preceding subscript c. The true statistics for both Table X H and X I are:

$$\tilde{V}_y = 1; \quad \tilde{V}_x = 1; \quad \tilde{c}_{xy} = .80; \quad \tilde{r}_{xy} = .80$$

Table X H (in which there is a slight error because the distributions have been assumed to terminate at $\pm 3.5\sigma$ yields:

$$V_{y_{\rm I}} = 1.08001; \qquad V_{x_{\rm I}} = 1.08001;$$

$$c_{x_{\rm I} y_{\rm I}} = .79728; \qquad r_{x_{\rm I} y_{\rm I}} = .73822$$

We note that there is negligible error in the covariance, but decided error in the variance and the correlation coefficient. In the case of a variable, such as a normal variable, having high order contact at both lower and upper ends, Sheppard's correction for the variance [6:49] is highly serviceable. Applying it we obtain,

$${}_cV_{y_{\rm I}} = .99668; \quad {}_cV_{x_{\rm I}} = .99668; \quad {}_cr_{x_{\rm I} y_{\rm I}} = .79994$$

Let us conclude that, having high order contact, Sheppard's correction to the variance will yield excellent approximations to true variances and correlations when the number of classes in each variable is six or greater.

No such excellent results are obtained with Table X I, where there are but three classes in each variable. We find (with a computational error because of terminating distributions at $\pm 3\sigma$) that:

$$V_{y_I} = 1.26924; \quad V_{x_I} = 1.26924; \quad c_{x_I y_I} = .78065$$

$$r_{x_I y_I} = .61505; \quad {}_cV_{y_I} = .93591; \quad {}_cV_{x_I} = .93591$$

$${}_c r_{x_I y_I} = .83411$$

These results suggest that there is only a small error in the covariance with very coarse grouping, but that the error in variance and correlation is serious in this instance and is not adequately allowed for by Sheppard's correction.

Corrections when class means are the variable: The class means given in Tables X H and X I are computed by formula [8:27]. Let these, which we will indicate by the subscript m, constitute the variables, and let the preceding subscript c′ indicate a correction for coarseness of grouping. Computation from Table X H yields:

$$V_{y_m} = .92267; \; V_{x_m} = .92267; \; c_{x_m y_m} = .68662; \; r_{x_m y_m} = .74417$$

From Table X I we obtain,

$$V_{y_m} = .73807; \; V_{x_m} = .73807; \; c_{x_m y_m} = .45395; \; r_{x_m y_m} = .61505$$

Clearly corrections to variance, covariance, and correlation are demanded if the problem is such as to make corrected results usable and interpretable.

Pearson (1913 Inf.), has shown that a correction based upon the correlation between the continuous variable, $\tilde{x}$, and its class means, x_m (and similarly for the second variable) is useful in correcting variance, covariance, and correlation.

This correlation, as proven by Kelley (1923) is

$$r_{\tilde{x}x_m} = \frac{\sigma_{x_m}}{\sigma_{\tilde{x}}} \quad \text{Correlation between a variate and the means of the classes into which it is recorded} \quad [10:91]$$

Thus

$$_{c'}V_{x_m} = \frac{V_{x_m}}{r^2_{\tilde{x}x_m}} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; [10:92]$$

This formula provides a "perfect" correction for the variance. It, in fact begs the question for all it asserts is that if one knows the true variance the correction to the class means variance is such as to produce the true variance. Also, as proven by Kelley (1923),

$$_{m}r_{x_m y_m} = \frac{r_{x_m y_m}}{r_{xx_m} r_{yy_m}} = \frac{r_{x_m y_m} \sigma_{\tilde{x}} \sigma_{\tilde{y}}}{\sigma_{x_m} \sigma_{y_m}} \quad \text{Coarseness of grouping correction to } r \text{ on account of class means} \quad [10:93]$$

If $\sigma_{\tilde{x}} = \sigma_{\tilde{y}} = 1$, as when a unit normal distribution is assumed, [10:93] becomes

$$_{m}r_{x_m y_m} = \frac{\Sigma x_m y_m}{N \; V_{x_m} \; V_{y_m}} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; [10:94]$$

The corrected covariance is,

$$_{c'}c_{x_m y_m} = \frac{r_{x_m y_m} V_{\tilde{x}} V_{\tilde{y}}}{\sigma_{x_m} \; \sigma_{y_m}} \quad \text{Coarseness of grouping correction to covariance on account of class means} \quad [10:95]$$

Applying these corrections to the data of Table X H, we obtain:

$_{c'}V_{y_m} = 1.0$; $_{c'}V_{x_m} = 1.0$; $_{c'}c_{x_m y_m} = .80654$; $_{c'}r_{x_m y_m} = .80654$

Applying these corrections to the data of Table X I we obtain:

$_{c'}V_{y_m} = 1.0$; $_{c'}V_{x_m} = 1.0$; $_{c'}c_{x_m y_m} = .83332$; $_{c'}r_{x_m y_m} = .83332$

Thus again we find correction for coarseness of grouping quite adequate in case we have seven classes, but not in case of three classes.

We conclude that *whether we employ class indexes or class means we have no clearly established procedure for corrections to variance, covariance, correlations, and regressions, for coarseness of grouping if the number of classes is 3, 4, or 5.* However, even with 3, 4, or 5 classes the correction formulas just given yield statistics which are closer to the true values than are the raw statistics.

Let us now classify the situations in which corrections for coarseness of grouping may be employed. If the problem is one of actual prediction and the predictor is only available, and in the future will be only available, in the coarse grouping no correction for coarseness of grouping in it should be made. If the predictand is only available in the coarse grouping and no utility attaches to a finer grouping, no correction for coarseness of grouping should be made in this variable. If the predictand, though only available in the coarse grouping, is the phenomenological expression of a noumenal continuum, knowledge of which would be of greater value than of the actual phenomena themselves, then a correction of the predictand for coarseness of grouping may be desirable. It is desirable if the relationship is substantial and the sample large so that no test of significance is needed, but it is not desirable if some sufficiently moot null hypothesis is postulated that a precise test is important. Finally, if both predictand and predictor are limited expressions of continuous variables, knowledge of the relationship between which answers an important theoretical issue, then corrections for coarseness of grouping in both variables is advisable, provided a need for a null hypothesis test is not primal.

Further corrections, especially for attenuation, for range, and for nonlinearity are considered in later chapters.

SECTION 14. CORRELATIONS AND OTHER STATISTICS OF SUMS AND DIFFERENCES

If

$$X_\alpha = w_1 X_1 + w_2 w_2 + \ldots + w_k X_k \qquad \text{[10:96a]}$$

A weighted sum

we may deal with deviations from means and write

$$x_\alpha = w_1 x_1 + w_2 x_2 + \ldots + w_k x_k$$

letting $\overset{k}{S}$ stand for a summation of k terms as i varies from 1 to k and letting $\overset{k(k-1)}{S}$ stand for a summation of $k(k-1)$ terms (i.e., the number of permutations of k things two at a time), as i and j vary from 1 to k under the restriction that $i \neq j$ it is readily derived that

$$V_\alpha = \overset{k}{S}\, w_i^2\, V_i + \overset{k(k-1)}{S} w_i w_j \sigma_i \sigma_j r_{ij} \qquad \text{[10:97]}$$

For a second weighted sum we use distinctive subscripts as follows:

$$X_\beta = w_a X_a + w_b X_b + \ldots + w_t X_t \qquad \text{[10:96b]}$$

The covariance between X_α and X_β is

$$c_{\alpha\beta} = \overset{kt}{S}\; w_i w_g \sigma_i \sigma_g r_{ig} \quad \left\{ \begin{array}{l} i = 1, 2 \ldots k \\ g = a, b \ldots t \end{array} \right. \qquad \text{[10:98]}$$

The correlation between these weighted sums is:

$$r_{\alpha\beta} = \frac{c_{\alpha\beta}}{\sigma_\alpha \sigma_\beta} \qquad \text{Correlation between a first and a second weighted sum} \qquad \text{[10:99]}$$

If all the measures which combined yield X_α are similar, as for example, would be the case if they are similar forms of some test, and if

the same holds for the measures which are combined to yield X_β we have approximately

$$\sigma_1 = \sigma_2 = \ldots \sigma_k$$

$$r_{12} = r_{13} = \ldots r_{1k} = \ldots r_{k-1,k}$$

$$\sigma_a = \sigma_b = \ldots \sigma_t$$

$$r_{ab} = r_{ac} = \ldots \quad r_{st} = \ldots r_{t-1,t}$$

We use $\bar{r}_{ij}$ to designate the average of the $k(k-1)/2$ correlations r_{12}, r_{13}, etc., We use $\bar{r}_{gh}$ to designate the average of the $t(t-1)/2$ correlations r_{ab}, r_{ac}, etc., and we use $\bar{r}_{ig}$ to designate the average of the kt correlations r_{1a}, r_{1b}, ... r_{kt}. We immediately obtain,

Correlation between sums, each consisting of equally weighted measures

$$r_{\alpha\beta} = \frac{kt\ \bar{r}_{ig}}{\sqrt{k+k(k-1)\bar{r}_{ij}}\sqrt{t+t(t-1)\bar{r}_{gh}}} \qquad [10:100]$$

As an approximation to [10:100] we may write [10:101]

Approximate correlation between sums when the measures entering each are similar, equally weighted, and equally variable

$$r_{\alpha\beta} = \frac{kt\ r_{1a}}{\sqrt{k+k(k-1)r_{12}}\ \sqrt{t+t(t-1)r_{ab}}} \qquad [10:101]$$

Precise equality of standard deviations may be brought about by normalizing measures of a series, or by ranking them, for then each variable will have rank scores from 1 to N.

Let us consider the case in which x_β consists of a single measure which we will now call x_0 and in which $x_\alpha = x_1 + x_2 + \ldots + x_k$, each of x_1, x_2, ... x_k being a series of ranks from 1 to

N. Then we can readily compute V_a from the data and obtain $\bar{\rho}_{ij}$ from [10:97] which in this case may be written

$$V_a = V_1 [k + k(k-1)\bar{\rho}_{ij}] = \frac{N^2-1}{12} [k + k(k-1)\bar{\rho}_{ij}] \qquad [10:102]$$

Yielding the average rank intercorrelation from the variance of a sum of ranks

Further

$$r_{0a} = \frac{k\ \sigma_1\ \bar{\rho}_{0i}}{\sigma_a} = \sqrt{\frac{N^2-1}{12}} \left(\frac{k\bar{\rho}_{0i}}{\sigma_a}\right) \qquad [10:103]$$

Yielding the average correlation of the separate parts of a sum (of measures which are ranks) with the criterion from the correlation of the sum with the criterion

Working with raw ranks we have

$$V_a = \frac{\Sigma\ X_a^2}{N} - \left[\frac{k(N-1)}{2}\right]^2 \qquad [10:104]$$

The variance of N measures, each being the sum of k rank scores

Combining [10:102] and [10:104] we obtain

$$\bar{\rho}_{ij} = \frac{2k\ (2N+1)}{(k-1)(N-1)} + \frac{12\ \Sigma\ X_a^2}{k(k-1)N(N^2-1)} \qquad [10:105]$$

Average intercorrelation between k series of N ranks each

The variance error of $\bar{\rho}_{ij}$ is

$$V(\bar{\rho}_{ij}) = \left[\frac{12}{(N^2-1)k(k-1)}\right]^2 V\ (V_a) \qquad [10:105a]$$

The following problem illustrates the use of this formula. Six judges, K, T, U, B, L, H rank according to merit twelve answers to a given problem as follows:

TABLE X J

RANKS GIVEN BY JUDGES

ANSWERS	K	T	U	B	L	H	X_α	X_α^2
A	1	5	7	10	2	5	30	900
B	2.5	6	4	6	3	9	30.5	930.25
C	2.5	3	1	4	1	2	13.5	182.25
D	4	2	2	11	8	3	30	900
E	5	12	3	1	4	10	35	1,225
F	6	1	8	2	5	1	23	529
G	7	11	10	8	12	4	52	2,704
H	8	9	5	7	6	11	46	2,116
I	9	4	9	12	7	6	47	2,209
J	10	7	11	5	9	8	50	2,500
K	11	10	12	9	10	12	64	4,096
L	12	8	6	3	11	7	47	2,209
								20,500.50

$$k = 6, \quad N = 12, \quad \Sigma X_\alpha^2 = 20,500.50$$

therefore, by [10:104], $\overline{\rho}_{ij} = .3241$.

We note that if k and t each approach infinity $r_{\alpha\beta}$ is simply a coefficient corrected for attenuation and [10:100] becomes [11:25].

If the weighted sum [10:96a] is simply $X_1 - X_2$, which we may call X_d, thus

$$X_d = X_1 - X_2$$

Then [10:97] reduces to

$$V_d = V_1 + V_2 - 2c_{12} = V_1 + V_2 - 2\sigma_1\sigma_2 r_{12} \qquad [10:106]$$

The variance of a difference

which is a formula of wide utility.

CHAPTER XI

FURTHER CORRELATION ISSUES

SECTION 1. THE VARIOUS CONSEQUENCES OF EMPLOYING SEMI-RELIABLE INITIAL MEASURES

Classic sampling theory has looked upon the N scores, X_1, in a sample as randomly selected from a parent population and has been concerned with the problem of estimating the characteristics of the population from the information revealed by the sample. The observed record or score of a single individual, or case, *a*, is considered to be an utterly trustworthy record. The concept of error attaches to X_{1a} only when X_{1a} is taken as an estimate of a population statistic, say of $\widetilde{M}_1$. Also a square difference $(X_{1a}-X_{1b})^2$ is exactly the squared difference of the difference between the scores of individuals *a* and *b* and as such has no error. Only when this squared difference is taken as evidence of the population variance of differences does the concept of error attach to it.

The concept of reliability of original measures is utterly other than this. The number of

kernels on an ear of corn from a certain stalk is not a perfect measure of the fruitfulness per ear of the stalk, if it has a second ear, because the number of kernels of this second ear may differ. The price of a wheat product at a certain date, relative to the price at a basic date, is not a perfect wheat index, for a different wheat product may yield a different index. The score of a pupil on a reading test is not a perfect measure of his reading ability for a second reading test given under similar conditions may reveal a different relative standing. The concept "reliability," or "reliability of measurement," concerns the phenomena of differences in the measurements obtained when the individual is doubly or multiply measured by means of instruments believed and intended to tap the same function. This concept has proven very useful in psychological and educational fields, where it has been widely employed, and we may anticipate a similar utility if widely used in other social fields and in the biological and physical sciences. In the physical sciences the concept of unreliability of original observations has a long and honorable history, being subsumed under the topic, "errors of observation," but the correlational consequences of these errors have not been pursued in this field to the extent that they have been in the field of mental measurement.

The concept only arises when two or more supposedly similar measures of the same thing are found to differ. Consider the case of an economist who has selected 40 products to incorporate into a cost of living index. These 40 items have been conceived by him as in some essential sense being expressions of a single phenomenon,—cost of living. His judgment has been employed in their selection and weighting, if differentially weighted. In combining the items

into a single final score, he has either thought that the record of each item was an expression of this common function-plus-chance, or of this common function-plus-a-non-chance-unique-function-plus-chance. Thus for the recorded score on item i of the index we have,

$x_i = x_c + x_u + e_i$, in which

x_c = the cost of living function

x_u = a non-chance, non-cost-of-living, function unique to item i, that is, x_u is not found (except as a matter of chance) in any of the other 39 items of the index.

e_i = an unaccountable, chance, or error influence that has affected x_i. It will correlate with no other function (except as a matter of chance).

The unique function, being unique, has the same correlational properties as the chance factor, so we combine them and designate the combination e_i. In brief, x_c is a real, or true, measure of the cost of living of which x_i is a fallible measure, and we write

$$x_i = c_i x_\omega + e_i$$

the c_i being a constant depending upon the particular units of measurement which have been employed. For a second item we have

$$x_j = c_j x_\omega + e_j$$

and similarly for further items. Any linear weighted combination of these 40 items can be written

$$x_1 = x_\omega + e_1 \qquad [11:01]$$

A fallible measure as a function of a true measure plus a factor operating like a chance factor

For some purposes we may find it convenient to write [11:01] thus

$$x_1 = c_1 x_\omega + e_1 \quad . \; . \; . \; . \; . \; . \quad [11:02]$$

which differs from [11:01] only in the units of measurement and not in the inherent relationship portrayed.

The expression for x_1 asserts that the economist conceives of his index as measuring nothing but cost of living plus irrelevant or chance factors. This attitude toward the total score, x_1, extends to the separate items, x_1, which have entered into it. He can therefore split the total score into halves. It is customary to represent the half scores by $x_{\frac{1}{2}}$ and $x_{\frac{I}{II}}$, but to simplify subscript notation we will here call the half scores x_3 and x_5.

$$\left.\begin{aligned} x_{\frac{1}{2}} &\equiv x_3 = .5x_\omega + e_3 \\ x_{\frac{I}{II}} &\equiv x_5 = .5x_\omega + e_5 \end{aligned}\right\} \quad \text{Scores on the halves of } x_1 \qquad [11:03]$$

In these equations the error factors are uncorrelated, being either strictly chance factors or having unique elements which behave like chance factors. Of course x_ω is uncorrelated with either chance factor. For the total measure we have

$$x_1 = x_3 + x_5 = x_\omega + e_1 \quad . \; . \; . \; . \; . \; . \; . \quad [11:01a]$$

If an x_2 measure is involved, we shall designate it

$$x_2 = x_4 + x_6 = x_\gamma + e_2 \quad . \; . \; . \; . \; . \; . \; . \quad [11:01b]$$

Also we shall designate a criterion measure

$$x_0 = x_\infty + e_0 \quad . \; . \; . \; . \; . \quad [11:04]$$

Let us estimate as fully as possible the statistics of x_∞ from a knowledge of x_1, x_3, and x_5. The variance of the differences between the half scores is

$$V(x_3 - x_5) = V_3 + V_5 - 2c_{35} = V_{e_3} + V_{e_5}$$
$$= 2V_{e_3} = V_{e_1} = V_{1.\omega} \qquad [11:05]$$

which is a function of the errors of measurement only. This gives us Rulon's (1939) formula for the standard error of measurement:

$$\sigma_{1.\omega} = \sigma_{x_3 - x_5} \qquad \text{Standard error of estimate when non-regressed } x_1 \text{ is taken as evidence of } x_\omega \qquad [11:06]$$

The notation $\sigma_{1.\omega}$ in this connection requires explanation. If one investigates the scatter diagram whose dimensions are x_ω and x_1, he can readily prove that the standard deviation of arrays of x_1 for fixed values of x_ω, the standard notation for which is $\sigma_{1.\omega}$, is exactly σ_{e_1}, [11:05] and [11:06]. Accordingly, $\sigma_{1.\omega}$ may correctly be used to indicate the standard error when x_1 is estimated from a knowledge of x_ω (a process actually never performed), or when x_ω is estimated by using non-regressed x_1.

The customary way to split a measure composed of many items into halves is to call the odd-numbered items one half and the even-numbered items the other half. This is frequently entirely satisfactory, but thought should be given to the matter, for if a different splitting yields halves which, in the judgment of the experimenter are more nearly comparable, then this different splitting should be employed. Surely in a cost-

of-living index all the semi-luxury items should not be assigned to the same half score, just as in a mental test all the hard items should not be concentrated in a single half. A priori considerations may suggest that some items have larger relative chance factors than others, and if so they should be judiciously distributed between the halves, etc. The writer (1942) would emphasize that *there should be full utilization of and dependence upon judgment in splitting a measure into halves, just as there has been such dependence upon it when drawing up the instrument in the first instance*. It is in fact desirable that the process of making the instrument be one of making comparable halves.

The suggestion has been made that a measure of $2t$ items can be split into halves in C_t^{2t} ways and that the best procedure would be to try out all these ways (when computing reliability coefficients) and find the average result. Theoretically this would seem to the writer sound only in the rare instance in which the devisor considered all items *(a)* equally excellent, *(b)* equally subject to chance, and *(c)* all functioning at the same level. Kuder and Richardson (1937, 1939) have devised formulas for the computation of reliability coefficients when these conditions hold. Their more elaborate and precise formulas scarcely constitute an economy of labor over the formulas for the reliability coefficient given in this section, though it is true that they void the perplexing question of how to split a test into halves. Their simplest formula for a test (*K-R* write in terms of mental test problems) the items of which are scored "right," or "wrong," is

$$r_1 = \frac{t}{t-1}\left(\frac{V_1 - t\,\overline{p}\,\overline{q}}{V_1}\right) \qquad \text{. . . . . [11:07]}$$

Kuder-Richardson short formula for the reliability coefficient

in which t is the number of items in the test, $\bar{p} = \frac{n}{t}$ = the mean item success score and $\bar{q} = 1-\bar{p}$. When the special conditions *(a)*, *(b)*, and *(c)* maintain, this reliability coefficient is the same as that given by [11:10], and it, rather surprisingly, approximates this answer for quite a range of common situations.

The standard error of estimate, given by Rulon's formula, is the statistic needed in order to attach the proper degree of confidence to an original measure. The numerical steps are straight-forward and also contribute to further statistics which are commonly needed. Simply record in four columns the values of x_3, x_5, (x_3-x_5), and x_1 $(=x_3+x_5)$, and compute the variance of each column. In addition to the essential variance error of estimate, $V(x_3-x_5)$, we can now readily obtain *C. S. Spearman's reliability coefficient*, which *is the correlation between similar measures*. Referring to the sum and difference formula for correlation we have

$$r_{\frac{1}{2}} \equiv r_{35} = \frac{V_1 - V(x_3 - x_5)}{4\,\sigma_3\,\sigma_5} \qquad [11{:}08]$$

The half-measure reliability via differences between half scores

We desire r_1, or as frequently designated r_{1I}, the correlation between x_1 and a postulated similar measure x_I. We shall obtain this by first deriving the Spearman-Brown step-up formula giving the reliability of a measure n times as long as the measure whose reliability has been experimentally determined. Usually one has computed r_{35} and desires r_1, in which case $n=2$, for x_1 is the sum of the two similar measures x_3 and x_5, the reliability of which is the computed value r_{35}. Let

$$x_n = x_3 + x_5 \; . \; . \; . + x_{2n+1}$$

be a measure which is the sum of n similar parts, and let x_N be a measure similar to x_n:

$$x_N = x_{III} + x_V + \ldots + x_{2N+I}$$

All parts of x_n and of x_N are similar.

$$V_n = V_N = nV_3 + (n^2-n)c_{35}$$

$$c_{nN} = n^2 c_{35}$$

$$r_n \equiv r_{nN} = \frac{n^2 c_{35}}{nV_3 + (n^2-n)c_{35}} = \frac{n\, r_{35}}{1+(n-1)r_{35}} \qquad [11:09]$$

Spearman-Brown step-up formula

Solving [11:09] for r_{35} we obtain the step down formula [11:09a]

$$r_{35} = \frac{r_n}{r_n + n(1-r_n)} \qquad [11:09a]$$

Solving [11:09] for n we obtain [11:09b]

$$n = \frac{\tilde{r}_n\,(1-r_{35})}{r_{35}(1-\tilde{r}_n)} \qquad [11:09b]$$

giving n when r_{35} is known and some reliability $\tilde{r}_n$ is desired. The variance error of n thus derived is

$$V_n = \frac{n^2}{r_{35}^2(1-r_{35})^2} V_{r_{35}} \doteqdot \frac{n^2(1+r_{35})^2}{r_{35}^2(N-2)} \qquad [11:09c]$$

For the step-up from the half measure reliability, [11:09] becomes,

$$r_1 = \frac{2r_{\frac{1}{2}}}{1+r_{\frac{1}{2}}} \equiv \frac{2r_{35}}{1+r_{35}} \qquad [11:10]$$

The variance error of r_n is consequent to the error in r_{35} and by the method of Chapter XIII, Section 8, we obtain Shen's (1924) formula for the standard error of r_n:

$$\sigma_{r_n} = \frac{n\,\sigma_{r_1}}{[1+(n-1)r_1]^2} \qquad [11:11]$$

The standard error of a stepped-up reliability coefficient

The standard error of r_1 is given by [10:46]: Utilizing this together with [11:10], we find that when the step-up is from the half score to the total score, then [11:11] becomes

$$\sigma_{r_1} = \frac{2(1-r_1)}{\sqrt{N-2}} \qquad [11:12]$$

Standard error of r_1 here derived via half scores. (Here N is number of cases in the sample.)

Utilizing relationships [11:05], [11:10], and [11:13],

$$V_1 = V_3 + V_5 + 2c_{35} = 2V_3(1+r_{35}) \qquad [11:13]$$

Variance of the sum of two similar measures

we obtain

$$V_{e_1} = V_1(1-r_1) \qquad [11:14]$$

Variance error of estimate as a function of the variance and reliability coefficient

Illustrations from the field of mental measurement of functions involving the reliability coefficient: A score on a psychological test is fallible and may be set equal to a true ability plus an error, as indicated in [11:01]. We clearly desire as much information as possible about the true ability, x_ω, having the observed test score, x_1. Dealing with gross scores, rather than deviation measures, we write,

$$X_1 = X_\omega + e_1 \qquad \text{Observed ability = true ability + an error} \qquad [11:01]$$

the e_1's being such that their mean for a large sample tends towards zero. This must be so for otherwise there is something in the e_1's which is not chance and this would then be a part of X_ω. Accordingly the mean of the X_1's is equal to the mean of the X_ω's within limits represented by the variance error M_1 for given values of X_ω, namely, $\tilde{V}_{e_1}/N$. This variance error is only $1/N$ as large as that attaching to the individual score so we may, unless N is very small, interpret individual scores on the basis that

$$M_\omega \doteq M_1 \qquad \text{The mean of true ability} \doteq \text{the mean of observed ability} \qquad [11:15]$$

Dealing with deviation scores, we can obtain the variance of both members of [11:01], utilize [11:14] and obtain

$$V_\omega = V_1 r_1 \qquad \text{An estimate of the variance of true ability} \qquad [11:16]$$

and

$$\sigma_\omega = \sigma_1 \sqrt{r_1} \qquad \text{An estimate of the standard deviation of true ability} \qquad [11:16a]$$

This and further relationships given involving "true ability" are strictly true in the population and approximately so in the sample. We note that the observed variability in a group tends to be greater than the true variability. The difference with instruments of low reliability may be so great as materially to change the picture of events. For a certain reading test, with reliability of about .36, an investigator found that some 47 per cent of fifth-grade pupils fell below the fourth-grade mean or above the sixth-grade mean, thus indicating serious misgrading or misclassification. Utilizing [11:16] and assuming a normal distribution, it is simple to compute the percentage which, in true ability, lies outside these same limits. The answer is 23,

only half the percentage given by the raw scores.

We can estimate an individual's X_ω from his X_1.

$$c_{1\omega} = \frac{1}{N}\Sigma x_1 x_\omega = \frac{1}{N}\Sigma(x_\omega + e_1)x_\omega = V_\omega \qquad [11:17]$$

Obtaining σ_ω from [11:16a] we find that

$$r_{1\omega} = \sqrt{r_1} \qquad \text{Correlation between a true and a fallible measure of the same function} \qquad [11:18]$$

$$\bar{x}_\omega = r_{1\omega}\frac{\sigma_\omega}{\sigma_1} x_1 = r_1 x_1 \qquad [11:19]$$

and

$$\bar{X}_\omega = r_1 X_1 + (1-r_1)\, M_1 \qquad \text{Regression of true ability upon a fallible measure of it} \qquad [11:20]$$

This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates,—one based upon the individual's observed score, X_1, and the other based upon the mean of the group to which he belongs, M_1. If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa. Suppose fourth-grade pupil A and fifth-grade pupil B each score 45 on a test having a reliability of .80 in each grade, and that the means and standard deviations for the grades are: M_4=40; σ_4=10; M_5=50; and σ_5=10. For pupil A we estimate his true ability thus:

$$\bar{X}_\infty = .80\ (45) + .20\ (40) = 44$$

For pupil B

$$\bar{X}_\omega = .80\ (45) + .20\ (50) = 46$$

This difference in outcome is certainly sound. We know two things about pupil A, the first

fact (X_1=45) suggests a true ability of 45, and the second fact (member of group whose mean = 40) suggests a true ability of 40. The best composite of his ability is 44, as given by [11:20]. Suppose for the single hour when tested pupil A had sat with the fifth grade, would we now use the fifth-grade mean and estimate his true ability as 46? Certainly not, for pupil A is still a fourth-grader. This group membership is not a whim, but a thing as definitely attached to pupil A as is his score 45.

The variance of the estimated true scores for the entire group is clearly

$$V_{\overline{\omega}} = r_1^2 \ V_1 \qquad \text{Variance of estimated true scores (see [11:22])} \qquad [11:21]$$

which we note is less than V_ω. The relationship $V_1 > V_\omega > V_{\overline{\omega}}$ is important.

The variance error of an estimated true score is given by the usual formula [10:09], which for this situation becomes

$$V_{\omega.1} = V_1(r_1 - r_1^2) \qquad \text{Variance error of estimate of a true score, } x_\omega \text{ from } x_1 \qquad [11:22]$$

This is less than the variance error of estimate given by [11:14]. The difference in the situations must be explicitly noted. When x_1 is taken as evidence of x_ω, the variance error of estimate is given by [11:14], but when x_1 is regressed so that $r_1 x_1$ is taken as evidence of x_ω, an improved estimate is gotten and the variance error of estimate is now given by [11:22]. If the mean and reliability for the group to which the tested person naturally belongs are known, it is always preferable to use the regressed score as the estimate of true ability. Since this best practice is infrequent practice, the pertinent variance error of estimate is usually that given by [11:14].

Formula [11:21] throws interesting light upon

the classification of individuals by fallible measures. Suppose upon a scholastic test we have fourth, fifth, and sixth-grade means of 40, 50, and 60, and that σ_{ω} for the fifth grade is 10. Assume a normal distribution of ability and a rule which demotes fifth-graders who score below 40 and promotes fifth-graders scoring above 60. If the reliability of the test is 1.00, we obtain the correct result of 16 per cent below 40 and 16 per cent above 60, and thus we reclassify 32 per cent of the pupils. If the test has a reliability of .50, we find with the aid of [11:16a] that $\sigma_1 = 14.14$. Employing raw scores, reference to a normal probability table informs us that we would now reclassify 48 per cent, which is an excessive number. If, however, we use regressed scores, $\bar{x}_{\omega} (= r_1 x_1)$, we have a distribution whose standard deviation is 7.07 (from [11:21]) and reclassify 16 per cent, which is a conservative number. It can also be shown, using volumes of a normal bivariate surface as tabled in Pearson (1931), that of the 48 per cent reclassified upon the basis of raw scores, 26/48 did not in truth fall beyond the limits set, and that of the 16 per cent reclassified upon the basis of regressed scores, 6/16 did not in truth fall beyond these limits. In short, the use of a fallible measure at its face value in connection with promotions, classification, etc., will lead to or create many misplacements, while the use of this same fallible measure properly regressed will create few misplacements. If we will but regress scores and compute standard errors of estimated true scores, we need not hesitate to use an instrument of low reliability.

The reliability of measures and issues involving two variables. Let x_{∞} be a true criterion, x_0 the fallible criterion of reliability r_0, r_{ω} a true predictor, and x_1 the fallible pre-

dictor of reliability r_1. We have the covariance

$$c_{01} = \frac{1}{N}\Sigma(x_\infty + e_0)(x_\omega + e_1)$$
$$= \frac{1}{N}(\Sigma x_\infty x_\omega + \Sigma x_\infty e_1 + \Sigma x_\omega e_0 + \Sigma e_0 e_1)$$

These last three summations are chance deviations from zero. We can set each = 0, an unbiased estimate of zero. We accordingly reach the following unbiased approximation:

$$c_{01} = c_{\infty 1} = c_{0\omega} = c_{\infty\omega} \quad . \; . \; . \; . \; . \; [11:23]$$

so that

$$r_{\infty 1} = \frac{c_{01}}{\sigma_\infty \sigma_1} = \frac{\sigma_0 \sigma_1 r_{01}}{\sigma_0 \sqrt{r_0}\sigma_1} = \frac{r_{01}}{\sqrt{r_0}} \qquad [11:24]$$

Correlation corrected for attenuation in one variable

By a similar derivation,

$$r_{\infty\omega} = \frac{r_{01}}{\sqrt{r_0}\sqrt{r_1}} \qquad [11:25]$$

C. S. Spearman's formula for correction for attenuation in both variables

If reliabilities are computed by finding the correlation between comparable half scores and stepping up, [11:10], the variance error of $r_{\infty 1}$ is given by formula [13:90], and the variance error of $r_{\omega\infty}$ by [13:88].

For the estimate of a true score we have, as given by [11:19],

$$\bar{x}_\infty = r_{\infty 1}\frac{\sigma_\infty}{\sigma_1}x_1 = r_{01}\frac{\sigma_0}{\sigma_1}x_1 \quad . \; . \; . \; . \; . \; . \; [11:26]$$

Since the estimate of x_0 from x_1 is

$$\bar{x}_0 = r_{01}\frac{\sigma_0}{\sigma_1}x_1$$

we observe that we reach the same answer in the two cases. However, this common answer is a more accurate estimate of the unknown true measure x_∞ than of the fallible measure x_0.

$$\sigma_{0.1} = \sigma_0 \sqrt{1-r_{01}^2}$$

as given by [10:49], and

$$\sigma_{\infty.1} = \sigma_\infty \sqrt{1-r_{\infty 1}^2} = \sigma_0 \sqrt{r_0 - r_{01}^2} \quad . \quad . \quad . \quad . \quad [11:27]$$

Standard error of estimate of a true criterion x_∞, from x_1

This is encouraging for it entitles us to believe, even if we do not know r_0, but do know that the criterion is fallible, that our actual estimates of the criterion are more trustworthy than indicated by $\sigma_0\sqrt{1-r_{01}^2}$. As illustration, we can cite the many studies reporting correlations with teachers' marks. In view of the fact that the reliability of teachers' marks seldom exceeds .75, that their average is in the neighborhood of .50 (lower for character ratings), and not infrequently are no higher than .30, we are entitled to attach considerably more significance to correlations with teachers' marks than has usually been done.

The significance of differences between fallible measures. The raw scores X_{1a} and X_{2a} of individual *a* upon two achievement measures are ordinarily in such units that the raw difference $X_{1a}-X_{2a}$ is not interpretable. If individual *a* belongs to a normally competitive group, a useful procedure is to express scores as standard scores using the mean and standard deviation of this group. Thus,

$$z_{1a} = (X_{1a} - M_1)/\sigma_1$$

and

$$z_{2a} = (X_{2a}-M_2)/\sigma_2$$

Then $z_{1a}-z_{2a}$ expresses the superiority of individual a in trait 1 to trait 2 relative to the group standard. A difference measure of this sort is doubly contaminated with error,—that in z_{1a} and that in z_{2a}. Counselors unfamiliar with this generally large error have foolishly trusted observed differences while other unduly conservative counselors have been prone to distrust all observed differences and place their trust in the general-mental-ability doctrine which considers the level of ability of a person to be the same in all intellectual functions. The precise handling of such within-the-individual differences is not difficult, but as one would expect, the standard error of the difference $z_{1a}-z_{2a}$ is not infrequently very large. If this same individual were tested with comparable forms of the two measures, his observed difference would be $z_{Ia}-z_{IIa}$. The correlation between such measures of difference for the group entire is the reliability of this difference in standard score measures. We may designate such a difference as $z_{1a}-z_{2a}$ with the symbol $d_{12.\omega\gamma}$, and read it "the difference between z_1 and z_2 for fixed values x_ω and x_γ," or "the difference between z_1 and z_2 which is independent of x_ω and x_γ." This is obviously so, for when the individual is the same, his true abilities x_ω and x_γ remain the same while the individual is being measured by X_1 or X_2.

$$d_{12} = z_1-z_2 \quad \text{. . . . . [11:28]}$$

$$V(d_{12}) = 2-2r_{12} \quad \text{. . . . . [11:28a]}$$

$$c(d_{12}d_{I\,II}) = V_\omega+V_\gamma-2c_{12} = r_1+r_2-2r_{12}. \quad \text{. . . . [11:29]}$$

$$r(d_{12}d_{\mathrm{I}\ \mathrm{II}}) = \frac{r_1 + r_2 - 2r_{12}}{2 - 2r_{12}} \qquad [11:29a]$$

Reliability of differences between the standard scores of an individual

$$d_{12.\omega\gamma} = z_{1.\omega\gamma} - z_{2.\omega\gamma} = e_1 - e_2 \quad \ldots\ldots \quad [11:30]$$

$$V(d_{12.\omega\gamma}) = 2 - r_1 - r_2 \qquad \text{Variance error of individual standard score difference} \quad [11:30a]$$

Formula [11:29a] shows that only in case r_{12} is zero or negative does the reliability of the within-the-individual difference compare favorably with that of the reliability of the measures employed in obtaining the difference.

$V(d_{12})$ can be analyzed into non-chance and chance independent parts, thus:

$$V(d_{12}) = [V(d_{12}) - V(d_{12.\omega\gamma})] + V(d_{12.\omega\gamma})$$

$$= V(d_{\omega\gamma}) + V(d_{12.\omega\gamma}) \quad \ldots\ldots \quad [11:31]$$

in which

$$d_{\omega\gamma} = \frac{z_\omega}{\sigma_{z_1}} - \frac{z_\gamma}{\sigma_{z_2}} = z_\omega - z_\gamma . \quad \ldots\ldots\ldots \quad [11:32]$$

$$V(d_{\omega\gamma}) = r_1 + r_2 - 2r_{12} \qquad \text{(See [11:29])} \qquad [11:32a]$$

Accordingly [11:31] may be written

$$2 - 2r_{1r} = (r_1 + r_2 - 2r_{12}) + (2 - r_1 - r_2) \qquad [11:31a]$$

$$\text{Total } V = \text{real } V \qquad + \text{chance } V$$

A precise variance ratio test involving $V(d_{\omega\gamma})$ and $V(d_{12.\omega\gamma})$ would be valuable, but is lacking

because the restrictions that have been placed upon the variables have been nonlinear in character.

Two other measures of difference are important. They are $d_{\omega/\omega,\gamma/\gamma}$ and $d_{\bar{\omega}\bar{\gamma}}$.

$$d_{\omega/\omega,\gamma/\gamma} = \frac{z_\omega}{\sigma_{z_\omega}} - \frac{z_\gamma}{\sigma_{z_\gamma}} = \frac{z_\omega}{\sqrt{r_1}} - \frac{z_\gamma}{\sqrt{r_2}} \qquad [11:33]$$

a non-individually observed difference, but one whose variance for the group can be computed. Its variance is

$$V(d_{\omega/\omega,\gamma/\gamma}) = 2-2r_{\omega\gamma} \qquad \text{A group measure of idiosyncrasy} \qquad [11:33a]$$

This measure of the extent to which the members of the group in their entirety are different (within themselves) in the two abilities would be of importance in a study of several groups which have had different types of tutelage. Further, if the variance error of $r_{\omega\gamma}$ as given by [13:88] is such that there is negligible probability that this variance is due to chance, it may then be appropriate to secure individual measures of difference, [11:28], or [11:34].

$$d_{\bar{\omega}\bar{\gamma}} = \frac{\bar{z}_\omega}{\sigma_{z_1}} - \frac{\bar{z}_\gamma}{\sigma_{z_2}} = r_1 z_1 - r_2 z_2 \qquad \text{Regressed measure of within-the-individual difference (see [11:37])} \qquad [11:34]$$

Writing $d_{\bar{\omega}\bar{\gamma}}$ thus:

$$d_{\bar{\omega}\bar{\gamma}} = r_1 z_1 - r_2 z_2 = (r_1 z_\omega - r_2 z_\gamma) + (r_1 e_1 - r_2 e_2)$$

we obtain an analysis of the variance into non-chance and chance portions.

$$V(d_{\bar{\omega}\bar{\gamma}}) = V(d_{r_1\omega,\, r_2\gamma}) + V(d_{\bar{\omega}\bar{\gamma}.\omega\gamma}) \qquad [11:34a]$$

$$r_1^2+r_2^2-2r_1r_2r_{12}=(r_1^3+r_2^3-2r_1r_2r_{12})$$
$$+(r_1^2-r_1^3+r_2^2-r_2^3) \qquad [11:\ 34b]$$

Of course the restrictions which have been imposed are nonlinear, so a variance ratio test would not be precise, but ordinarily we can secure serviceable information by dealing with critical ratios. The individual $d_{\bar{\omega}\bar{\gamma}}/\sigma(d_{\bar{\omega}\bar{\gamma}.\omega\gamma})$ is a critical ratio. The mean square of such may be compared with the mean square of $d_{12}/\sigma(d_{12.\omega\gamma})$ in order to ascertain which measure, $d_{\bar{\omega}\bar{\gamma}}$ or d_{12}, is the more efficient.

$$\frac{V(d_{12})}{V(d_{12.\omega\gamma})}=\frac{2-2r_{12}}{2-r_1-r_2} \quad \ldots\ldots\ldots \quad [11:35]$$

$$\frac{V(d_{\bar{\omega}\bar{\gamma}})}{V(d_{\bar{\omega}\bar{\gamma}.\omega\gamma})}=\frac{r_1^r+r_2^2-2r_1r_2r_{12}}{r_1^2-r_1^3+r_2^2-r_2^3} \quad \ldots \quad [11:36]$$

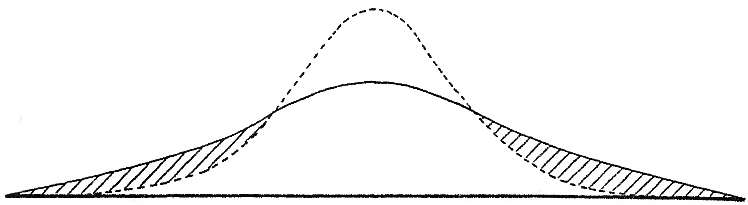
Substitution of numerical values in these equations shows that in case reliabilities are greater than .50 the regressed, or $d_{\bar{\omega}\bar{\gamma}}$, measure of difference is more trustworthy than the non-regressed, or d_{12}, measure.

The proportion of cases whose idiosyncrasies (within individual differences) are revealed by fallible measures. We will consider this problem when the observed measure of difference is d_{12} and second when $d_{\bar{\omega}\bar{\gamma}}$. Let us assume that all distributions of differences are normal. The ratio $\sigma(d_{12.\bar{\omega}\bar{\gamma}})/\sigma(d_{12})$ is the ratio of the standard deviation of differences due to chance to the standard deviation of the observed differences. As this ratio becomes small an increasing number of the observed differences are

not artifacts.

Chart XI I illustrates the situation. The dash-curve is the distribution of differences

CHART XI I



attributable to chance and the full-line curve is that of the observed differences, so that the shaded portion is the proportion of differences in excess of the chance proportion. Table XI A, herewith, (see Kelley, 1923), gives this pro-

TABLE XI A

NON-CHANCE DIFFERENCES FOR DIFFERENT RATIOS

$\frac{\sigma \text{ DUE TO CHANCE}}{\sigma \text{ OBSERVED}}$	PROPORTION OF DIFFERENCES IN EXCESS OF THE CHANCE PROPORTION	$\frac{\sigma \text{ DUE TO CHANCE}}{\sigma \text{ OBSERVED}}$	PROPORTION OF DIFFERENCES IN EXCESS OF THE CHANCE PROPORTION
.02	.950	.50	.323
.05	.888	.55	.281
.10	.798	.60	.242
.15	.719	.65	.205
.20	.647	.70	.171
.25	.582	.75	.138
.30	.522	.80	.108
.35	.467	.85	.078
.40	.415	.90	.051
.45	.367	.95	.025
		.99	.005

portion for different values of the ratio. Chart

XI I approximately represents the situation found in connection with Stanford Achievement Test measures of Arithmetic Computation and of Arithmetic Reasoning, for 96 eighth-grade pupils.

Arithmetic Computation score = x_1 and has a reliability of .669.

Arithmetic Reasoning score = x_2 and has a reliability of .825.

$$\sigma(d_{12.\omega\gamma}) = \sqrt{2-r_1-r_2} = .711$$

$$\sigma(d_{12}) = \sqrt{2-2r_{12}} = 1.246$$

Ratio = .571, equivalent to 26 per cent of cases having differences in excess of chance.

$$\sigma(d_{\bar{\omega}\bar{\gamma}.\omega\gamma}) = \sqrt{r_1^2-r_1^3+r_2^2-r_1^3} = .517$$

$$\sigma(d_{\bar{\omega}\bar{\gamma}}) = \sqrt{r_1^2+r_2^2-2r_1r_2r_{12}} = .939$$

Ratio = .551, equivalent to 27 per cent of cases having differences in excess of chance.

For these same data the reliability of d_{12}, as given by [11:29a], is .674, and the reliability of $d_{\bar{\omega}\bar{\gamma}}$, as given by [11:37] is .697.

$$r(d_{\bar{\omega}\bar{\gamma}}) = \frac{r_1^3+r_2^3-2r_1r_2r_{12}}{r_1^2+r_2^2-2r_1r_2r_{12}} \quad \text{(See [11:34])} \qquad [11:37]$$

Reliability and other issues involving three or more unequally reliable measures of the same thing

A typical situation would exist when several judges separately rate, or score, the same set of objects, be they people, cattle, art products, or what not. The rating by each judge is upon

some commonly defined scale like "impulsiveness of individuals," "general excellence of beef stock," "beauty of art sample," etc. Since the impulsiveness of a person can scarcely be better defined than as equal to the average rating, or score, given by an adequate number of those competent to rate, it is appropriate and useful to assert that the judges are rating the same thing, though they may be unequally expert in doing it. Similarly for many other situations. We may thus write for the scores as given by the k judges

$$x_1 = c_1 x_\omega + e_1; \quad x_2 = c_2 x_\omega + e_2; \ldots \quad x_k = c_k x_\omega + e_k$$

wherein the c's change from judge to judge and depend upon the respective scales used, wherein x_ω is the true score attaching to the object scores; and wherein the e's are errors of judgment, uncorrelated with each other or with the true scores.

When we have three judges it follows, as shown by Shen* (1925), that the reliability of a single judge, say judge number 1 is

$$r_1 = \frac{r_{12} r_{13}}{r_{23}} \quad \text{Triad formula for reliability} \quad [11:38]$$

The variance error of r_1 thus determined is

$$V_{r_1} = \frac{r_1^2}{N-2} \left[4r_1 + \frac{2}{r_1} + \frac{1}{r_{23}^2} + (1-2r_1)\right.$$

$$\left.\left(\frac{1}{r_{12}^2} + \frac{1}{r_{13}^2}\right) - 5\right] \quad [11:39]$$

* Formula [11:39] herewith differs slightly from Shen's formula in which there is a typographical error.

If the number of measures of the same thing is some number k, greater than three, we first express each set of measures as standard scores, thus $z_1 = (X_1-M_1)/c_1$; $z_2 = (X_2-M_2)/c_2$; ... $z_k = (X_k-M_k)/c_k$. We let $s = z_1 + z_2 + \ldots + z_k$, and let $s_1 = s - z_1$. Since the correlation between z_1 and s_1 corrected for attenuation is equal to 1.00, we have the reliability of the first set of measures as derived by Shen [11.40]

$$r_1 = \frac{r_{1s_1}^2}{r_{s_1}} = \frac{(k-2)\bar{r}_{1i}^2}{k\bar{r}_{ij}-2\bar{r}_{1i}} \qquad [11:40]$$

Reliability from the intercorrelations between unequally excellent measures of the same thing

In this formula $\bar{r}_{ij}$ is the average of the $k(k-1)/2$ intercorrelations between the k measures, and $\bar{r}_{1i}$ is the average of the $(k-1)$ correlations between z_1 and the $(k-1)$ remaining measures.

Frequently $\bar{r}_{ij}$ can be gotten by [10:105] and r_{1i} by [10:103] or by these formulas modified to be appropriate to standard score variables.

If we have four measures of the same thing, but unequally reliable, it is simple to prove that the three following tetrads, two of which are independent, equal zero.

$$\begin{aligned} t_{1234} &= r_{12}r_{34} - r_{13}r_{24} = 0 \\ t_{1342} &= r_{13}r_{24} - r_{14}r_{23} = 0 \\ t_{1243} &= r_{12}r_{34} - r_{14}r_{23} = 0 \end{aligned} \qquad [11:41]$$

Value of tetrads when a single general factor accounts for all intercorrelations

The variance error of t_{1234} as derived by Kelley (1928) is

$$V(t_{1234}) = \frac{1}{N-2}[r_{12}^2 + r_{13}^2 + r_{24}^2 + r_{34}^2 + 2r_{12}r_{14}r_{23}r_{34}$$

$$+ 2r_{13}r_{14}r_{23}r_{24} - 2r_{12}r_{13}r_{23} - 2r_{12}r_{14}r_{24}$$

$$- 2r_{13}r_{14}r_{34} - 2r_{23}r_{24}r_{34} + t_{1234}^2(r_{12}^2 + r_{13}^2$$

$$+ r_{14}^2 + r_{23}^2 + r_{24}^2 + r_{34}^2 - 4)] \qquad [11:42]$$

Thus, in the case of four variables, we may say that they may be conceived as due to a general factor plus four specific factors when

$$r_{12}r_{34} = r_{13}r_{24} = r_{14}r_{23}$$

Kelley has shown that the pentad function following equals zero when five variables may be thought of as consequent to two general factors plus specific factors:

$$f_{12345} = r_{12}r_{13}r_{24}r_{35}r_{45} + r_{14}r_{15}r_{23}r_{24}r_{35} + r_{12}r_{14}r_{25}r_{34}r_{35}$$

$$+ r_{13}r_{15}r_{24}r_{25}r_{34} + r_{12}r_{15}r_{23}r_{34}r_{45} + r_{13}r_{14}r_{23}r_{25}r_{45}$$

$$- r_{12}r_{14}r_{23}r_{35}r_{45} - r_{13}r_{15}r_{23}r_{24}r_{45} - r_{12}r_{15}r_{24}r_{34}r_{35}$$

$$- r_{13}r_{14}r_{24}r_{25}r_{35} - r_{12}r_{13}r_{25}r_{34}r_{45} - r_{14}r_{15}r_{23}r_{25}r_{34}$$

[11:43]

The computation of the variance error of the function is rather laborious. The relationships immediately preceding and still other general relationships connected with "factor analysis" and the minimal necessary number of independent variables underlying a given set of correlations can be well expressed in the notation of matrix algebra.

The weighting factor which is to be attached to a variable because of its unreliability. Let the standard scores z_1, z_2, z_3, . . . , each be measures of z (and of no other true measure) and have chance errors e_1, e_2, e_3, . . . , whose variances are unequal. To allow for different units of measurement we write:

$$z_1 = a_1 z_\omega + e_1, \text{ and } V_1 = 1 = a_1^2 V_\omega + (1-r_1)$$

$$z_2 = a_2 z_\omega + e_2, \text{ and } V_2 = 1 = a_2^2 V_\omega + (1-r_2)$$

$$z_3 = a_3 z_\omega + e_3, \text{ and } V_3 = 1 = a_3^2 V_\omega + (1-r_3) \text{ etc.}$$

(V_ω as here defined being different from V_ω as defined by [11:16] and [11:01].)
From the variances we note that

$$\frac{a_1}{\sqrt{r_1}} = \frac{a_2}{\sqrt{r_2}} = \frac{a_3}{\sqrt{r_3}} = \text{etc.}$$

so that we may now redefine z_ω and write the z's thus:

$$z_1 = \sqrt{r_1}\, z_\omega + e_1$$

$$z_2 = \sqrt{r_2}\, z_\omega + e_2$$

Taking the variance of z_1 we have, $1 = r_1 V_\omega + (1-r_1)$, so $V_\omega = 1$.

$$\frac{z_1}{\sqrt{r_1}} = z_\omega + \frac{e_1}{\sqrt{r_1}}$$

$$\frac{z_2}{\sqrt{r_2}} = z_\omega + \frac{e_2}{\sqrt{r_2}}$$

etc.

The left-hand members are unbiased estimates of z_ω. By [13:19] the best combination of them will be that given by weighting them inversely as their variance errors. The variance error of

$z_1/\sqrt{r_1}$ (equal to that of $e_1/\sqrt{r_1}$), as given by [11:14] is $(1-r_1)/r_1$. The inverse of this is the proper weighting factor of $z_1/\sqrt{r_1}$, so that the best estimate of z_ω is:

$$\frac{\dfrac{\sqrt{r_1}}{1-r_1}z_1 + \dfrac{\sqrt{r_2}}{1-r_2}z_2 + \dfrac{\sqrt{r_3}}{1-r_3}z_3 + \ldots}{\dfrac{r_1}{1-r_1} + \dfrac{r_2}{1-r_2} + \dfrac{r_3}{1-r_3} + \ldots} \quad [11:44]$$

Otherwise expressed, the ratios of weighting factors to allow for differences in reliabilities as otherwise derived by Kelley (1927, pp. 211-213) are

$$\frac{\sqrt{r_1}}{1-r_1} : \frac{\sqrt{r_2}}{1-r_2} : \frac{\sqrt{r_3}}{1-r_3} : \text{etc.} \quad \ldots\ldots\ldots \quad [11:45]$$

Of course these weights do not overtly modify or supplant the weights in a multiple regression equation, but lacking regression equation weights these weighting factors may confidently be expected to be beneficial. Table XI B herewith provides these weights for different reliability coefficients.

TABLE XI B

WEIGHTING FACTORS CONSEQUENT TO VALUES OF THE RELIABILITY COEFFICIENT

r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$
.01	.101	.26	.689	.51	1.46	.76	3.63
.02	.144	.27	.712	.52	1.50	.77	3.82
.03	.179	.28	.735-	.53	1.55-	.78	4.01
.04	.208	.29	.758	.54	1.60	.79	4.23
.05	.235+	.30	.782	.55	1.65-	.80	4.47

TABLE XI B
(CONTINUED)

r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$	r_1	$\frac{\sqrt{r_1}}{1-r_1}$
.06	.261	.31	.807	.56	1.70	.81	4.74
.07	.284	.32	.832	.57	1.76	.82	5.03
.08	.307	.33	.857	.58	1.81	.83	5.36
.09	.330	.34	.883	.59	1.87	.84	5.73
.10	.351	.35	.910	.60	1.94	.85	6.15-
.11	.373	.36	.938	.61	2.00	.86	6.62
.12	.394	.37	.966	.62	2.07	.87	7.17
.13	.414	.38	.994	.63	2.15-	.88	7.82
.14	.435+	.39	1.024	.64	2.22	.89	8.58
.15	.456	.40	1.054	.65	2.30	.90	9.49
.16	.476	.41	1.085+	.66	2.39	.91	10.60
.17	.497	.42	1.117	.67	2.48	.92	11.99
.18	.517	.43	1.150+	.68	2.58	.93	13.78
.19	.538	.44	1.185-	.69	2.68	.94	16.16
.20	.559	.45	1.220	.70	2.79	.95	19.49
.21	.580	.46	1.256	.71	2.91	.96	24.49
.22	.601	.47	1.294	.72	3.03	.97	32.83
.23	.623	.48	1.332	.73	3.16	.98	49.50
.24	.645-	.49	1.373	.74	3.31	.99	99.50
.25	.667	.50	1.414	.75	3.46		

SECTION 2. THE EFFECT OF VARIABILITY IN RANGE UPON CORRELATION

The effect of differences in range of talent examined upon reliability coefficients. Consider a test with scores, say, from 0 to 50, yielding a mean of about 25 when given to the group to which it is best adapted. If given to a markedly inferior group the scores will tend to cluster around 0, and if to a markedly superior group they will tend to cluster around 50, and if given to a group so heterogeneous as to include in-

ferior, average, and superior, the reliability scatter diagram involving scores X_1 and X_2 upon two similar forms of the test will tend to have lanceolate contour lines. This is evidence of malfunctioning of the instrument at both low and high levels. If the malfunctioning is at one of these levels only, pear-shaped contour lines will be found. *In connection with reliability the first observation to make is of the reliability scatter diagram in order to ascertain the range of effectiveness of the instrument.* The lanceolate contour lines indicate poor functioning of the instrument, but this is not revealed by the reliability coefficient, which in this case is high and accordingly very misleading.

If the instrument functions equally well throughout the range of talent tested, then the variability of differences (X_1-X_I) will tend to be the same at all levels. The variability of (X_1-X_I) is a variability at right angles to the major axis of the contour ellipses in the (X_1, X_I) scatter diagram and can be estimated by inspection of the scatter diagram, or it can be computed for different levels, a subject's level being determined by $(X_1+X_I)/2$. For an instrument of uniform excellence we have

$$V(x_1-x_I) = V_1+V_I-2\sigma_1\sigma_I r_{1I} \doteqdot 2V_1(1-r_1)$$

thus

$$V_1(1-r_1) = \text{a constant} \qquad [11:46]$$

is a situation which maintains throughout the range of equal effectiveness of the instrument. This immediately provides us with a means of estimating the reliability for one range of talent knowing it for a different range. We will use lower-case letters to indicate a narrower range

and capital letters a wider range. We have:

$$v_1(1-r_1) = V_1(1-R_1) \quad \text{Relationship between reliability and variability in the case of an instrument of uniform merit} \quad [11:46a]$$

Introducing the variability of true scores, this relationship becomes:

$$\frac{v_\omega}{V_\omega} = \frac{r_1(1-R_1)}{R_1(1-r_1)} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [11:47]$$

We may note that $v_1(1-r_1)$ is the variance of (x_1-x_ω) so that equality, for different ranges, of the variance of the differences (x_1-x_I) is equivalent to the equality, for different ranges, of the variance of the difference between observed scores and true scores (see Kelley, 1921).

Table XI C, herewith, illustrates how closely relationship [11:46a] is maintained in the case of one instrument designed to be effective from the second to the ninth grades.*

TABLE XI C

DATA UPON STANFORD SPELLING TEST, FORMS A AND B
Number of cases approximately 165 per school grade

SCHOOL GRADES	MEANS	VARIANCES	ACTUAL RELIABILITY COEFFICIENTS	RELIABILITY FROM TOTAL USING (11:46a)
2	26.0	339.	.88	.81
3	48.8	390.	.87	.83
4	71.6	374.	.87	.82
5	96.8	594.	.90	.89
6	117.8	569.	.87	.88
7	139.4	454.	.82	.86
8	160.4	515.	.76	.87
9	172.2	465.	.90	.86
2 to 9 inclusive	104.1	2901.	.9774	

* Kelley, Ruch, and Terman, STANFORD ACHIEVEMENT TEST MANUAL OF DIRECTIONS, 1925 and revised, 1927, Tables 2 and 4.

The differences between the actual and the estimated grade reliability coefficients are somewhat greater than may reasonably be attributed to chance. However, lacking specific narrow range determinations, formula [11:46a] should be used to estimate the narrow-range reliability knowing the wide-range value.

For any problem the reliability coefficient to be used is that which is appropriate to the normal competitive group implied in the problem. In school matters the normal competitive group is the school grade and certainly not the group given by combining grades 2 to 9. Even to report a reliability coefficient for a noncompetitive wide-range group is misinformative to all except the statistically expert.

When an individual may be considered as either a member of a narrow or of a wide-range group we ask the question of whether the connection in which we study him makes any difference. We have two formula [11:20] estimates of his ability:

$$\bar{X}_\omega = r_1 X_1 + (1-r_1)m_1, \quad \text{using narrow-range statistics;}$$

$$\bar{X}'_\omega = R_1 X_1 + (1-R_1)M_1, \quad \text{using wide-range statistics.}$$

Formula [11:22] provides us with the variance errors of these estimates. Utilizing [11:47] we find that the ratio of these variance errors is

$$\frac{v_{\omega.1}}{V_{\omega.1}} = \frac{v_1(r_1-r_1^2)}{V_1(R_1-R_1^2)} = \frac{r_1}{R_1} < 1$$

showing that more reliable results are always obtained by using the narrow-range group. In short, *if the individual may be treated as a member of a number of groups, the most reliable*

results will be obtained by treating him as a member of the most homogeneous of these, which, of course, is the group for which the reliability coefficient is the smallest.

The effect of curtailment in range of a first variable upon the reliability coefficient of a second variable. If a reliability coefficient is computed for a range of talent which has been restricted in a known manner upon the basis of another variable, it obviously is, in general, an inaccurate measure of the reliability when no such restriction is present. For example, the reliability of a test, X_2, computed for those who have attained a certain passing mark in X_1 is, in case X_1 and X_2 are correlated, smaller than the reliability of X_2 for a group consisting of those who have failed as well as passed, X_1. As deduced from a formula given by Davis (1944), the reliability, R_2, in the wide range, knowing that in the narrow range, r_2, knowing the correlation, r_{12}, in the narrow range, and knowing V_1/v_1, the ratio of the unrestricted variance in X_1 to the restricted variance, is

$$R_2 = \frac{r_2 + r_{12}^2\left(\frac{V_1}{v_1} - 1\right)}{1 + r_{12}^2\left(\frac{V_1}{v_1} - 1\right)} \quad \ldots\ldots \quad [11:48]$$

The effect of curtailing the range of one of the variables entering into a correlation problem. Let the variables have the correlation R_{12} and let it further be true that the arrays are homoscedastic so that $V_{1.2}$ is constant for successive X_2 values. If there is an imposed alteration in the X_2 values so that one or more of the arrays of X_1 values are removed entirely, or if a randomly chosen number of cases is withdrawn from any one of these arrays, it will not change

either $V_{1.2}$ or $B_{1.2}$. Thus, letting lower-case letters stand for the situation after the imposed curtailment in the X_2 measures and a consequential curtailment in the X_1 measures, we have:

$$v_{1.2} = V_{1.2};\ v_1(1-r_{12}^2) = V_1(1-R_{12}^2) \quad . \ . \quad [11:49]$$

$$b_{12} = B_{12};\ r_{12}^2 \frac{v_1}{v_2} = R_{12}^2 \frac{V_1}{V_2} \quad . \ . \ . \ . \ . \quad [11:50]$$

so that

$$\frac{r_{12}^2}{v_2(1-r_{12}^2)} = \frac{R_{12}^2}{V_2(1-R_{12}^2)} \quad \text{Relationship when a } v_2/V_2 \text{ relationship is imposed} \quad [11:51]$$

Of course, when selection is on the basis of the second variable, $V_{2.1} \neq V_{2.1}$ and $b_{21} \neq B_{21}$. Though there has been alteration in the variances of both variables, only the v_2/V_2 ratio is material in estimating the change in the correlation coefficient. Solving [11:51] for r_{12}^2, we have, as derived by Pearson (1902 Inf.),

$$r_{12}^2 = \frac{R_{12}^2 \dfrac{v_2}{V_2}}{1-R_{12}^2 + R_{12}^2 \dfrac{v_2}{V_2}} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ [11:52]$$

Table XI D gives the change in correlation with different curtailments (see Kelley, 1923).

TABLE XI D

RELATIONSHIP BETWEEN CORRELATION AND AN IMPOSED CURTAILMENT OF ONE OF THE VARIABLES

σ_2 / Σ_2	IF r=.1 THEN R=	r=.2 R =	r=.3 R =	r=.4 R =	r=.6 R =	r=.8 R =	r=.95 R =
.75	.133	.263	.387	.503	.707	.872	.971
.50	.197	.378	.532	.658	.832	.936	.987
.25	.373	.632	.783	.868	.949	.983	.997
.10	.709	.898	.953	.975	.991	.997	.9995

An interesting possible use of [11:51] lies in the study of biological phenomena. If, for some form of life, in one geographical region two characters have the constants v_1, v_2, r_{12}, and in a second distant region the constants V_1, V_2, R_{12}, and if it is believed that the life in one region is a migrant and variable form of that in the other, then the two characters X_1 and X_2 may be investigated to see which is the more causal in causing the selection. If $r_{12}^2/[v_2(1-r_{12}^2)]$ is more nearly equal to $R_{12}^2/[V_2(1-R_{12}^2)]$ than $r_{12}^2/[v_1(1-r_{12}^2)]$ is to $R_{12}^2/[V_1(1-R_{12}^2)]$ then the indication is that X_2 is more closely connected with the cause of selection than is X_1.

If the curtailment is an arbitrary cutting off in the case of one variable of a portion of a distribution which is believed to be approximately normal when uncurtailed, the ratio v_2/V_2 can be readily determined, knowing the portion which has been deleted. For example, let us say that Test X_2 is employed as an admission-to-training-in-the-Air-Corps instrument; that the form of distribution on this test of all applicants is approximately normal; that arrays are essentially homoscedastic; and that the lowest 25 per cent of those taking the test are not admitted. For those admitted, the correlation of X_2 with attained flying efficiency is r_{12}. What reasonably would have been the correlation had all the applicants been admitted to training? We find the ratio of v_2/V_2 by [8:29], introduce the result into [11:52] and find the R_{12} which is equivalent to r_{12}. Had r_{12} been .50 the answer would be $R_{12} = .62$.*

The effect of double selection upon correlation. This problem was attacked by Pearson

* This procedure extensively used by Carl Brown, in Robert M. Yerkes, ED., PSYCHOLOGICAL EXAMINING in the U.S.Army, Nat. Acad. of Sci., Vol. 15, 1921.

(1908) in the case of normal bivariate correlation. The explicit solution is provided in Formula [11:53], derived by Vern James.*

$$r_{12}=\frac{-(1-R_{12}^2)\Sigma_1\Sigma_2+\sqrt{(1-R_{12}^2)^2V_1V_2+4v_1v_2R_{12}^2}}{2\,\sigma_1\,\sigma_2\,R_{12}} \qquad [11:53]$$

The specific situation to which this formula applies requires that *(a)* the distribution prior to selection, represented by capital letters, be a normal bivariate distribution and *(b)* that the distribution after selection, represented by lower-case letters, be a normal bivariate distribution.

Formula [11:53] may be written

$$\frac{r_{12}^2}{v_1v_2(1-r_{12}^2)}=\frac{R_{12}^2}{V_1V_2(1-R_{12}^2)} \quad \ldots\ldots \quad [11:54]$$

or again

$$\frac{r_{12}}{\sigma_{1.2}\,\sigma_{2.1}}=\frac{R_{12}}{\Sigma_{1.2}\,\Sigma_{2.1}} \quad \ldots\ldots\ldots\ldots \quad [11:55]$$

revealing an invariant relationship which may well be important in the study of natural selection.

* A class paper submitted to the writer by Vern James, February, 1925. James made the following happy substitutions: He let Pearson's $(1-t_1^2)=x$; Pearson's $(1-t_2^2)=y$; Pearson's $\Sigma_1/\sigma_1=a$; Pearson's $\Sigma_2/\sigma_2=b$.

SECTION 3. THREE-VARIABLE MULTIPLE CORRELATION

The three-variable problem will be approached in several ways because it involves practically all of the issues inherent in the general, or k-variable, problem and is algebraically much simpler than the general multiple-correlation problem.

Let X_0, the criterion or predictand, be a measure which it is desired to estimate from a knowledge of X_1 and X_2, the independent variables or predictors. There is so great a simplification when the regressions of X_0 are linear that it is desirable to transform X_1 and X_2 into new variables yielding linear relationships with X_0, if a preliminary study of the X_0, X_1 scatter diagram, and/or of the X_0, X_2 scatter diagram, reveals demonstrable nonlinearity. Though nonlinear regression of X_1 or X_2 upon X_0 is not a major detriment, the writer has frequently found that a single transformation of the X_0 variable will frequently rectify the regression of X_0 upon a number of other variables, X_1, X_2, X_3, etc.,—a result otherwise accomplished only by making transformations in all of the predictor variables.

In the following treatment we assume that the regression of X_0 upon X_1 and upon X_2 is approximately linear, and that an equation of the type

$$\overline{X}_0 = a + b_1X_1 + b_2X_2 \quad \text{3-variable multiple regression equation} \quad [11:56]$$

$$\overline{x}_0 = b_1 x_1 + b_2x_2 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad [11:56a]$$

will constitute a sound basis for estimating X_0. A more explicit notation for the regression coefficients is $b_{01.2}$,—read "the regression of X_0 upon X_1 for X_2 constant",—and $b_{02.1}$. Where there is no ambiguity we shall use the shorter notation for both b and β (see [11:57]) regres-

sion coefficients.

Let standard z_0, z_1, and z_2 scores be as defined by [8:23]. Then [11.56] is represented by an equation of the type

$$\bar{z}_0 = \beta_1 z_1 + \beta_2 z_2 \qquad \text{3-variable } \beta \text{ regression equation} \qquad [11:57]$$

We have

$$b_1 = \beta_1\frac{\sigma_0}{\sigma_1}, \text{ and } b_2 = \beta_2\frac{\sigma_0}{\sigma_2} \quad \ldots\ldots [11:58]$$

$$a = M_0 - b_1 M_1 - b_2 M_2 \quad \ldots\ldots\ldots\ldots [11:59]$$

When the standard score z_0 is estimated by means of [11:57] the variance error of estimate, $k^2_{0.12}$, is as follows:

$$k^2_{0.12} = \frac{1}{N}\Sigma(z_0 - \beta_1 z_1 - \beta_2 z_2)^2 \quad \ldots\ldots [11:60]$$

$$= 1 + \beta_1^2 + \beta_2^2 - 2\beta_1 r_{01} - 2\beta_2 r_{02} + 2\beta_1\beta_2 r_{12}$$

It is desired so to determine the β's that $k^2_{0.12}$ shall be a minimum. We may, without loss of generality, write $\beta_1 = \beta_2\gamma$ and introduce γ into equation [11:60]. We obtain,

$$k^2_{0.12} = [\beta_2^2\gamma^2 - 2\beta_2\gamma(r_{01} - \beta_2 r_{12}) + (r_{01} - \beta_2 r_{12})^2]$$

$$+ 1 + \beta_2^2 - 2\beta_2 r_{02} - (r_{01} - \beta_2 r_{12})^2$$

To make $k^2_{0.12}$ a minimum γ must be given such a value as to make the [] term, which is a perfect square, equal to zero. Then

$$k^2_{0.12} = 1 + \beta_2^2 - 2\beta_2 r_{02} - r_{01}^2 + 2\beta_2 r_{01} r_{12} - \beta_2^2 r_{12}^2 = [\beta_2^2(1 - r_{12}^2)$$

$$-2\beta_2(r_{02} - r_{01}r_{12}) + \frac{(r_{02} - r_{01}r_{12})^2}{1 - r_{12}^2}] - \frac{(r_{02} - r_{01}r_{12})^2}{1 - r_{12}^2} + 1 - r_{01}^2$$

This is a minimum when the [] term, which is a perfect square, is equal to zero, thus

$$\beta_2 = \frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2} \qquad [11:61]$$

β, or standard score, multiple regression coefficient (3 variables)

$$\beta_1 = \frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2} \quad \ldots\ldots\ldots\ldots \quad [11:61a]$$

$$k_{0.12}^2 = \frac{1 - r_{01}^2 - r_{02}^2 - r_{12}^2 + 2r_{01}r_{02}r_{12}}{1 - r_{12}^2} \qquad [11:62]$$

Variance error of estimate of standard score criterion (3 variables)

The correlation between z_0 and $\overline{z}_0$ is the multiple correlation coefficient and is always positive. Since there may be various estimates of z_0, a more explicit notation than $z_{\overline{0}}$ is needed. We shall use $z_{0\Delta12}$, which may be read "that part of z_0 which is dependent upon z_1 and z_2." The subscript Δ may be thought of as standing for the word "dependent," just as in $z_{0.12}$,—"that part of z_0 which is independent of z_1 and z_2,"—the subscript dot may be thought of as the dot to the letter *i* in the word "independent." The correlation between z_0 and the best linear combination of z_1 and z_2 is a multiple correlation. We shall designate it $r_{0\Delta12}$. Other notations are $r_{0(12)}$, $R_{0(12)}$, and $r_{0.12}$, but this last, which the writer has employed earlier, will be avoided as incongruous with the "independent of" concept which, as here used, attaches to the subscript dot. For the correlation between z_0 and that part of z_0 which is independent of z_1 and z_2, we shall, as in the preceding derivation, use the notation $k_{0.12}$. Though $r_{0.12}$ would be an entirely logical notation for this, it

will not be used in this text because of its earlier use in the opposite sense. Utilizing the principle inherent in [10:16] and [10:17],

$$V(z_0) = V(z_{0\Delta 12}) + V(z_{0.12})$$

$$1 \quad = (1-k^2_{0.12}) + k^2_{0.12} \equiv r^2_{0\Delta 12} + k^2_{0.12} \quad [11:63]$$

Some computational procedures yield $k^2_{0.12}$, in which case we compute $r^2_{0\Delta 12}$ from the equation

$$r^2_{0\Delta 12} = 1 - k^2_{0.12} \quad . \quad . \quad . \quad . \quad . \quad . \quad [11:64]$$

Other computational procedures yield the β regression coefficients, in which case it is simple to compute $r_{0\bar{0}}$, which is identical with $r_{0\Delta 12}$, thus, when we note that $\sigma(z_{0\Delta 12}) = r_{0\Delta 12}$,

$$r_{0\Delta 12} = r_{0(0\Delta 12)} = \frac{\Sigma z_0(\beta_1 z_1 + \beta_2 z_2)}{N\ \sigma(z_{0\Delta 12})} = \frac{\beta_1 r_{01} + \beta_2 r_{02}}{r_{0\Delta 12}}$$

$$r^2_{0\Delta 12} = \beta_1 r_{01} + \beta_2 r_{02} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [11:65]$$

Computational formula for $r^2_{0\Delta 12}$ knowing β regression coefficients

The variance error of z_0, as estimated by means of [11:57], is $k^2_{0.12}$, and of X_0, as estimated by means of [11:56], is

$$V_{0.12} = V_0 k^2_{0.12} = V_0(1-r^2_{0\Delta 12}) \quad . \quad . \quad . \quad . \quad . \quad [11:66]$$

Variance error of estimate,—3 variables

Corresponding to the analysis of variance equation [11:63] is the following which involves raw X_0 scores:

$$V_0 = V_{0\Delta 12} = V_{0.12}$$

$$V_0 = V_0 r^2_{0\Delta 12} + V_0(1 - r^2_{0\Delta 12}) \quad \text{Analysis of the variance of } X_0 \text{ scores} \qquad [11:67]$$

$$N-1 = 2 + (N-3) \quad \text{Degrees of freedom equation} \qquad [11:68]$$

That the residual variance $V_{0.12}$ has $(N-3)$ degrees of freedom is obvious when we note that the following three linear restrictions, and none other, have been imposed upon the X_0 scores:

$$NM_0 = \Sigma X_0; \quad Nc_{01} = \Sigma(X_0 - M_0)x_1; \quad Nc_{02} = \Sigma(X_0 - M_0)x_2$$

The nonlinear restriction $NV_0 = \Sigma(X_0 - M_0)^2$ is not involved in a, b_1, or b_2. b_1 may be written

$$b_1 = \beta_1 \frac{\sigma_0}{\sigma_1} = \frac{\dfrac{c_{01}}{\sigma_1} - \dfrac{c_{02}}{\sigma_2} r_{12}}{(1 - r^2_{12})\sigma_1}$$

a function not involving σ_0, or V_0. Similarly for b_2. We can get the number of degrees of freedom of $V_{0\Delta 12}$ by subtracting $(N-3)$ from $(N-1)$, or by noting that the regression equation [11:56] has two degrees of freedom due to the two parameters b_1 and b_2, for M_0, the additional parameter involved in a, has already established the degrees of freedom of V_0 as $(N-1)$, not N.

The variance ratio,

$$F_{2, N-3} = \frac{V_{0\Delta 12}/2}{V_{0.12}/(N-3)} = \frac{r^2_{0\Delta 12}/2}{(1 - r^2_{0\Delta 12})/(N-3)} \qquad [11:69]$$

provides a test for the significance of $(r^2_{0\Delta 12} - 0)$.

In the special case in which $r_{12} = 0$,—a situ-

ation which can frequently be brought about by an appropriate design of an experiment,—we note that all of the correlations between x_1, x_2, and $x_{0.12}$ are equal to zero, so that we have:

$$x_0 = x_{0\Delta 12} + x_{0.12} = b_1x_1 + b_2x_2 + x_{0.12}$$

$$V_0 = b_1^2V_1 + b_2^2V_2 + V_{0.12} \quad \text{Analysis of } V_0 \text{ when } x_1 \text{ and } x_2 \text{ are independent, 3 variables} \qquad [11:70]$$

$$N-1 = 1 + 1 + (N-3) \quad \text{Degrees of freedom equation} \qquad [11:71]$$

The variance ratios,

$$F_{1,N-3} = \frac{b_1^2 \ V_1/1}{V_{0.12}/(N-3)}, \text{ and}$$

$$F_{1,N-3} = \frac{b_2^2 \ V_2/1}{V_{0.12}/(N-3)} \quad \dots\dots\dots \qquad [11:72]$$

Variance ratios when predictors are uncorrelated provide tests for the significance of (b_1^2-0) and of (b_2^2-0). If we desire a test for the deviation of b_1 from $\tilde{b}_1$, a value not zero, we have,

$$F_{1,N-3} = \frac{(b_1-\tilde{b}_1)^2 \ V_1/1}{V_{0.12}/(N-3)} \quad \text{(predictors independent) (cf [11:78])} \qquad [11:73]$$

The standard score regression coefficients, β_1 and β_2 may be interpreted in connection with *partial correlation*, which *is the correlation between parts of variables*. If $z_{0.2}$ is that part of z_0 that is independent of z_2 it is given by $z_{0.2} = z_0 - r_{02}z_2$. Its variance is $k_{02}^2 = 1-r_{02}^2$. Similarly $z_{1.2} = z_1 - r_{12}z_2$, and the variance of

$z_{1.2}$ is $k_{12}^2 = 1-r_{12}^2$. The correlation between $z_{0.2}$ and $z_{1.2}$ is written $r_{01.2}$ and is read "the correlation between variables 0 and 1 independent of the effect of variable 2," or "the correlation between variables 0 and 1 for variable 2 constant."

$$r_{01.2} = \frac{\Sigma z_{0.2} z_{1.2}}{N k_{0.2} k_{1.2}} = \frac{r_{01} - r_{02} r_{12}}{\sqrt{1-r_{02}^2}\ \sqrt{1-r_{12}^2}} \qquad [11:74]$$

Partial correlation coefficient,- 3 variables

Utilizing [11:61a] we obtain

$$\beta_1 \equiv \beta_{01.2} = r_{01.2} \frac{k_{02}}{k_{12}} \quad \ldots\ldots\ldots \quad [11:75]$$

Relationship between regression and partial correlation coefficients,-standard scores

$$b_1 \equiv b_{01.2} = r_{01.2} \frac{\sigma_{0.2}}{\sigma_{1.2}} \quad \ldots\ldots\ldots \quad [11:76]$$

Relation between regression and partial correlation coefficients

Yule (1907) and Fisher (1923-24) have shown the similarity in the properties of and the distribution of the partial correlation coefficient to those of a total correlation coefficient. The number of degrees of freedom in the partial correlation coefficient is less than in the total correlation coefficient by the number of secondary subscripts (those after the point) in the partial correlation coefficient.

If $x_{1.2}$ were used to estimate $x_{0.2}$, the regression equation would be

$$\bar{x}_{0.2} = r_{01.2} \frac{\sigma_{0.2}}{\sigma_{1.2}} x_{1.2} = b_{01.2}\, x_{1.2} \equiv b_1\, x_{1.2} \qquad [11:77]$$

Thus we note that the proper weight to attach to

x_1, in [11:56a], in estimating x_0 is identically that which would attach to that part of x_1 which is independent of x_2 when estimating that part of x_0 which is independent of x_2. Similarly for the other independent variable.

Reference to [11:77] informs us that $b_{01.2}$ is a regression coefficient in a two-variable problem in which the variables are $x_{0.2}$ and $x_{1.2}$, each having N-2 degrees of freedom, so that a precise variance ratio test is given by [10:39], which becomes

$$F_{1,N-3} = \frac{[(b_{01.2}-\tilde{b}_{01.2})^2 V_{1.2}](N-3)}{V_{0.2}(1-r^2_{01.2})}$$

Since $V_{1.2} = V_1 k^2_{12}$ and $V_{0.2}(1-r^2_{01.2}) = V_{0.12} = V_0 k^2_{0.12}$

$$F_{1,N-3} = \frac{[(b_{01.2}-\tilde{b}_{01.2})^2 V_1 k^2_{12}\ (N-3)}{V_0 \quad k^2_{0.12}} \qquad [11:78]$$

F to test $(b_{01.2}-\tilde{b}_{01.2})$. See [12:48]

The parent population must here be conceived as one in which the identical x_1 and x_2 values as paired in this sample are repeated upon all successive samplings. Under these conditions of sampling there will result a succession of values of $b_{01.2}$. It is the variability of these that is tested by [11:78]. Since this variance ratio has one degree of freedom in the numerator variance, we may write

$$V(b_{01.2}) = \frac{V_0 k^2_{0.12}}{V_1 k^2_{12}(N-3)} \quad \text{Variance error of regression coefficient, three variables} \quad [11:79]$$

Determinantal expression of multiple correlation relationships: We herewith give these relationships for the three-variable problem and state, without proof, that they also hold for the general case of k variables, if expressed in connection with a major determinant, of the type Δ herewith, but expanded to include the added variables. For the three-variable case we have:

$$\Delta = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{01} & 1 & r_{12} \\ r_{02} & r_{12} & 1 \end{vmatrix} \qquad \text{Major correlation determinant, 3 variables} \qquad [11:80]$$

$$\Delta_{01} = \begin{vmatrix} r_{01} & r_{12} \\ r_{02} & 1 \end{vmatrix} \qquad \text{Minor obtained by deleting the 0 row and 1 column}$$

Other minors are similarly designated and A, with subscripts, stands for a co-factor, which is a minor with a sign factor attached.

$$1 - r^2_{0\Delta 12} = k^2_{0.12} = \frac{\Delta}{\Delta_{00}} \qquad [11:81]$$

Multiple alienation coefficient, or variance error of estimated standard scores

$$\beta_{01.2} = \beta_1 = \frac{\Delta_{01}}{\Delta_{00}} = \frac{-A_{01}}{A_{00}} \quad \ldots\ldots\ldots \quad [11:82]$$

$$\beta_{02.1} = \beta_2 = \frac{-\Delta_{02}}{\Delta_{00}} = \frac{-A_{02}}{A_{00}} \quad \ldots\ldots\ldots \quad [11:82a]$$

$$r_{01.2} = \sqrt{\beta_{01.2}\beta_{10.2}} = \sqrt{\frac{\Delta_{01}\Delta_{10}}{\Delta_{00}\Delta_{11}}} = \sqrt{\frac{A_{01}A_{10}}{A_{00}A_{11}}} \qquad [11:83]$$

$\beta_{01.2}$ and $\beta_{10.2}$ have the same sign, and the sign that attaches to the radical is this sign. These two β's are conjugate regression coefficients.

SECTION 4. NONLINEAR REGRESSION

Notation illustrated in connection with a three-variable linear regression situation. The theoretical issues and the illustrative problem will refer to a quadric or second-degree parabola, but the method can be immediately extended to higher-degree parabolic regression. It is, in fact, so slight a modification of multiple-regression procedure as to call for few new concepts. A new notation will be found convenient. By illustrating this for a three-variable linear multiple-regression problem, the similarity between that problem and parabolic regression will be made apparent. In connection with the three-variable linear-regression problem there are slightly different procedures dependent upon whether the task set is to compute the constants of [11:56], [11.56a], or [11.57]. The major determinant pertinent to the determination of the constants in [11:57] will be called A, or Δ, that pertinent to [11:56a] δ, that pertinent to [11:56] d, or D, depending upon whether means or summations are employed. We employ p to designate a product moment involving x, or deviation, scores, P a product moment involving X, or gross scores, and Σ a summation of gross scores. Σ, P, and p will each, in this three-variable problem, have three subscripts, the first indicating the power to which the first variable (X_0 or x_0) is raised, the second the power to which the second variable (X_1 or x_1) is raised, and the third the power to which the third variable (X_2 or x_2) is raised. Thus

$$p_{ijk} = \frac{\Sigma x_0^i x_1^j x_2^k}{N}$$

$$P_{ijk} = \frac{\Sigma X_0^i X_1^j X_2^k}{N}$$

$$\Sigma ijk = \Sigma X_0^i X_1^j X_2^k$$

Such substitutions as $\Sigma_{001} = NM_2$; $P_{100} = M_0$; $p_{110} = c_{01}$; $p_{020} = V_1$; etc., can be made whenever desired.

$$A = \Delta = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{01} & 1 & r_{12} \\ r_{02} & r_{12} & 1 \end{vmatrix} \quad \ldots\ldots\ldots\ldots \quad [11:80]$$

$$\delta = \begin{vmatrix} p_{200} & p_{110} & p_{101} \\ p_{110} & p_{020} & p_{011} \\ p_{101} & p_{011} & p_{002} \end{vmatrix} = \begin{vmatrix} V_0 & c_{01} & c_{02} \\ c_{01} & V_1 & c_{12} \\ c_{02} & c_{12} & V_2 \end{vmatrix} \quad [11:84]$$

$$d = \begin{vmatrix} P_{200} & P_{100} & P_{110} & P_{101} \\ P_{100} & P_{000} & P_{010} & P_{001} \\ P_{110} & P_{010} & P_{020} & P_{011} \\ P_{101} & P_{001} & P_{011} & P_{002} \end{vmatrix} = \begin{vmatrix} P_{200} & M_0 & P_{110} & P_{101} \\ M_0 & 1 & M_1 & M_2 \\ P_{110} & M_1 & P_{020} & P_{011} \\ P_{101} & M_2 & P_{011} & P_{002} \end{vmatrix} \quad [11:85]$$

$$D = \begin{vmatrix} \Sigma_{200} & \Sigma_{100} & \Sigma_{110} & \Sigma_{101} \\ \Sigma_{100} & \Sigma_{000} & \Sigma_{010} & \Sigma_{001} \\ \Sigma_{110} & \Sigma_{010} & \Sigma_{020} & \Sigma_{011} \\ \Sigma_{101} & \Sigma_{001} & \Sigma_{011} & \Sigma_{002} \end{vmatrix} \quad \text{in which } \Sigma_{000} = N \quad [11:86]$$

The complete solution, starting with Δ, is as indicated in Section 3. The solution starting

starting with δ or d will be obvious from that based upon D which is given herewith.

When using subscripts we will call the first row of D the 0-row; the second row the a-row; the third row the 1-row; and the fourth row the 2-row; and similarly for the columns. Thus D_{0a} is the cofactor of the element Σ_{100}. Note that the determination of the sign factor, which is simple in any case, must be as though the numbering were 0, 1, 2, 3 and not as here used, 0, a, 1, 2.

The various statistics, most of which are unnecessary if the regression equation only is desired, are

$$M_0 = \frac{\Sigma_{100}}{N}; \quad M_1 = \frac{\Sigma_{010}}{N}; \quad M_2 = \frac{\Sigma_{001}}{N} \quad . \ . \ . \ [11:87]$$

$$V_0 = \frac{\Sigma_{200}}{N} - M_0^2; \quad V_1 = \frac{\Sigma_{020}}{N} - M_1^2; \quad V_2 = \frac{\Sigma_{002}}{N} - M_2^2 \quad [11:88]$$

$$r_{01} = \frac{\frac{\Sigma_{110}}{N} - M_0 M_1}{\sigma_0 \sigma_1}; \quad r_{02} = \frac{\frac{\Sigma_{101}}{N} - M_0 M_2}{\sigma_0 \sigma_2};$$

$$. \ . \ . \ [11:89]$$

$$r_{12} = \frac{\frac{\Sigma_{011}}{N} - M_1 M_2}{\sigma_1 \ \sigma_2}$$

$$a = \frac{-D_{0a}}{D_{00}}; \ b_1 = \frac{-D_{01}}{D_{00}}; \ b_2 = \frac{-D_{02}}{D_{00}} \ [11:90], [11:91], [11:92]$$

$$V_{0.12} = V_0 k_{0.12}^2 = V_0(1 - r_{0\Delta 12}^2) = \frac{D}{N D_{00}} \ \text{Variance error of estimate} \ [11:93]$$

$$r^2_{0\Delta 12} = 1 - \frac{D}{NV_0 D_{00}} \qquad \text{Squared multiple correlation} \qquad [11:94]$$

$$r_{01.2} = \sqrt{b_{01.2}b_{10.2}} = \sqrt{\frac{D^2_{01}}{D_{00}D_{11}}} \; . \; . \; . \; . \; . [11:95]$$

Partial correlation coefficient
(sign of $r_{01.2}$ is that of $b_{01.2}$)

$$r_{02.1} = \sqrt{\frac{D^2_{12}}{D_{00}D_{22}}}\;;\; r_{12.0} = \sqrt{\frac{D^2_{12}}{D_{11}D_{22}}}$$

Though the proof is omitted, it can be stated that the preceding formulas of this Section hold for problems of any number of variables by simply adding the necessary secondary subscripts to $V_{0.12}$, $r_{0\Delta 12}$, $r_{01.2}$, etc.

Parabolic regression. If the third variable is simply the second squared, the regression equation is

$$X_{0\Delta 1,1^2} = A + BX_1 + CX_1^2 \qquad \text{Second-degree parabolic regression} \qquad [11:96]$$

The linear estimate, $\overline{X}_0$, may be written

$$X_{0\Delta 1} = a + bX_1 \qquad \text{Linear regression} \qquad [11:97]$$

In the case of [11:96] the correlation obtained may be designated $r_{0\Delta 1,1^2}$ and called a second-degree parabolic correlation coefficient.

If we substitute X_1^2 for X_2 in [11:56] the major determinant [11:86] then becomes

$$D = \begin{vmatrix} \Sigma_{20} & \Sigma_{10} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{10} & \Sigma_{00} & \Sigma_{01} & \Sigma_{02} \\ \Sigma_{11} & \Sigma_{01} & \Sigma_{02} & \Sigma_{03} \\ \Sigma_{12} & \Sigma_{02} & \Sigma_{03} & \Sigma_{04} \end{vmatrix} \quad \ldots\ldots\ldots\ldots [11:98]$$

and thence the procedure is as before. The calculation of the variance error, of the parabolic correlation, and of the regression constants A, B, and C follows the pattern of [11:93], [11:94], [11:90], [11:91], and [11:92], but the variances of B and C do not follow the pattern of b_1 and b_2 [11:79]. C could be written $b_{01^2.1}$, and read, "the regression of X_0 upon X_1^2 for fixed values of X_1." However, when X_1 is fixed of course X_1^2 is also fixed. We are unable to determine the variance of $b_{02.1}$ without fixing $b_{01.2}$, (see Chapter XII), but now a different and more powerful procedure is open to us.

When X_0 is estimated from a knowledge of X_1, we have [11:97] and we can analyze V_0 into independent parts,

$$V_0 = V_{0\Delta 1} + V_{0.1} \quad \ldots\ldots\ldots\ldots [11:99]$$

$$N-1 = 1 + (N-2) \quad \text{Corresponding d.o.f. equation} \quad [11:100]$$

$V_{0\Delta 1}$ is the variance of X_0 dependent upon X_1 and equals $V_0 r_{01}^2$. $V_{0\Delta 1}$ is the variance of the points on the linear regression line and $V_{0.1}$ is the variance of the divergences from these points.

In the case of second-degree parabolic regression we have [11:96] and we can analyze V_0 into the following independent parts:

$$V_0 = V_{0\Delta 1,1^2} + V_{0.1,1^2} \quad \ldots\ldots\ldots\ldots [11:101]$$

$$N-1 = 2 + (N-3) \quad \text{Corresponding d.o.f. equation} \quad [11:102]$$

$V_{0\Delta 1,1^2}$ is the variance of the points on the second-degree parabolic regression line and $V_{0.1,1^2}$ is the variance of the divergences from these points.

Also $V_{0.1}$ of [11:99] can be analyzed into independent parts. A part which can be estimated from X_1^2 but not from X_1 and a part which cannot be so estimated:

$$V_{0.1} = V_{(0.1)\Delta(1^2.1)} + V_{(0.1).(1^2.1)}$$

$$= V_{(0.1)\Delta(1^2.1)} + V_{0.1,1^2} \quad \ldots\ldots [11:103]$$

Substituting in [11:99] we have

$$V_0 = V_{0\Delta 1} = V_{(0.1)\Delta(1^2.1)} + V_{0.1,1^2}$$

$$= V_0 r_{01}^2 + V_0(r_{0\Delta 1,1^2}^2 - r_{01}^2) + V_0(1 - r_{0\Delta 1,1^2}^2) \qquad [11:104]$$

$$N-1 = 1 + 1 + (N-3) \quad \text{d.o.f. equation} \qquad [11:105]$$

$V_{0\Delta 1}$ = the variance of the points on the linear regression line.

$V_{(0.1)\Delta(1^2.1)}$ = the variance of the differences between the points on the second-degree line and the points on the linear line.

$V_{0.1,1^2}$ = the residual variance, or that of the divergences from the second-degree regression line.

The variance ratio test of the need of a linear regression line, —b in [11:97] $\neq 0$, is

$$F_{1(N-2)} = \frac{r_{01}^2(N-2)}{1 - r_{01}^2} \quad F \text{ to test need of at least linear regression} \qquad [11:106]$$

Again, we have

$$F_{1(N-3)} = \frac{(r^2_{0\Delta 1,1^2} - r^2_{01})(N-3)}{1 - r^2_{0\Delta 1,1^2}} \qquad [11:107]$$

F to test need of at least second-degree regression

Similarly,

$$F_{1(N-4)} = \frac{(r^2_{0\Delta 1,1^2 1^3} - r^2_{0\Delta 1,1^2})(N-4)}{1 - r^2_{0\Delta 1,1^2 1^3}} \qquad [11:108]$$

F to test need of at least third-degree regression

And so forth to any desired degree of parabolic regression that one may desire to test.

If the X_1 variable is grouped into k classes, a parabola of degree $(k-1)$ will pass exactly through the means of the arrays so that a parabola of higher degree will give no better fit. The F test will show this, but prior to this, if successively higher-degree parabolas have been employed, a P-from-F > .5 may have been obtained. This would indicate that the degree of the parabola used was too high,—a better fit being gotten than was to be expected from the size of the sample dealt with. When successive tests are made, one should stop with the parabola of such degree that P from F at this point is approximately equal to .5.

Another procedure, leading to the same outcome and frequently less time-consuming, is to compute ϵ^2, the unbiased correlation ratio squared, and terminate the testing of higher and higher-degree parabolas when the r^2 obtained is negligibly less than ϵ^2. Practical considerations might assert that an r^2 of, say, .800, was negligibly less than an ϵ^2 of .802, even though the sample was so large that P from F at this point < .5.

SECTION 5. THE CORRELATION RATIO

This method is of specific value in testing the adequacy of linear and other regression lines. Also, it is a very general method in that it applies when one variable is quantitative and the other qualitative.

In the case of linear regression, the predictand, X_0, can be expressed as the sum of the prediction, $X_{0\Delta 1}$, and an error of estimate, $X_{0.1}$, The four equations following have already been established:

$$X_0 = X_{0\Delta 1} + X_{0.1}$$

$$V_0 = V_{0\Delta 1} + V_{0.1} \qquad \text{Analysis of variance equation}$$

$$N-1 = 1 + (N-2) \qquad \text{D.o.f. equation}$$

$$r^2_{01} \equiv r^2_{0\Delta 1} = V_{0\Delta 1}/V_0$$

For quadric regression we have:

$$X_0 = X_{0\Delta 1,1^2} + X_{0.1,1^2}$$

$$V_0 = V_{0\Delta 1,1^2} + V_{0.1,1^2}$$

$$N-1 = 2 + (N-3)$$

$$r^2_{0\Delta 1,1^2} = V_{0\Delta 1,1^2}/V_0$$

Similarly for higher parabolic regression up to that of k-1 degree, at which point a perfect fit is attained, that is, the regression line passes exactly through the means of the k successive arrays. At this point we have:

$$X_0 = X_{0\Delta 1,1^2\ldots 1^{k-1}} + X_{0.1,1^2\ldots 1^{k-1}}$$

$$V = V_{0\Delta 1,1^2\ldots 1^{k-1}} + V_{0.1,1^2\ldots 1^{k-1}}$$

$$N-1 = (k-1) + (N-k)$$

$$r_{0\Delta 1,1^2\ldots 1^{k-1}} = V_{0\Delta 1,1^2\ldots 1^{k-1}} / V_0$$

The variance $V_{0\Delta 1,1^2\ldots 1^{k-1}}$ is simply that of the means of the X_0's for the k successive arrays of X_0's. Designating these means M_{0i}, in which $i=1, 2,\ldots k$, and the squared correlation coefficient in this case of perfect fit η^2_{01}, we have

$$\eta^2_{01} = \frac{V(M_{0i})}{V_0} \equiv 1 - \frac{V_{0.1,1^2\ldots 1^{k-1}}}{V_0} \qquad [11:109]$$

The raw correlation ratio squared

To obtain $V(M_{0i})$ compute the k means of the arrays, weight each by the number of cases in the array, and compute the variance.

The numerical values, if there are any, that attach to the k class indexes of the X_1's have not entered into this computation so no ordered or quantitative relationship between the X_1 classes enters into this measure. Accordingly the correlation ratio is applicable when a quantitative relationship between the X_1 classes is unknown.

Also, of course, if there is a known quantitative relationship between the X_1 classes, the employment of η^2 uses but a part of the available information.

It is not to be expected that any regression line of less degree than $k-1$ will exactly hit the means of all the arrays. Accordingly $V_{0\Delta 1}$,

$V_{0\Delta 1,1^2}$, $V_{0\Delta 1,1^2 1^3} \ldots V_{0\Delta 1,1^2 \ldots 1^{k-2}}$ constitute increasing series whose limit is $V(M_{0|})$. Also $r^2_{0\Delta 1}$, $r^2_{0\Delta 1,1^2}, \ldots r^2_{0\Delta 1,1^2 \ldots 1^{k-2}}$ constitute an increasing series whose limit is η^2_{01}. The difference between η^2_{01} and $r^2_{0\Delta 1.1^2 \ldots 1^j}$ is a measure of the adequacy of the j-degree regression line to account for the relationship between X_0 and X_1.

However, when we bear in mind that the relationship that we seek to discover is that maintaining in the population and not that in the sample we must believe that a regression line that fits the sample perfectly overstates the intimacy of the relationship in the population. We should therefore discount, or shrink, both r^2 and η^2 to get estimates of the population values before we compare them. Both r^2 and η^2 are equal to expressions of the sort

$$1 - \frac{V\text{ (errors of estimate)}}{V_0}$$

A similar expression using estimates of the population variances will yield the desired shrunken values of r^2 and η^2, which we label ${}_s r^2$ and ϵ^2. For estimated population variances we have:

$$\tilde{\tilde{V}}_0 = V_0 \frac{N}{N-1} \quad \ldots \ldots \ldots \ldots \quad \text{See [6:09]}$$

$$\tilde{\tilde{V}}_{0.1} = V_{0.1} \frac{N}{N-2} \quad \ldots \ldots \ldots \ldots \quad [11:110]$$

$$\tilde{\tilde{V}}_{0.1,1^2} = V_{0.1,1^2} \frac{N}{N-3} \quad \ldots \ldots \ldots \quad [11:111]$$

$$\tilde{\tilde{V}}_{0.1,1^2\ldots 1^j} = V_{0.1,1^2\ldots 1^j}\frac{N}{N-j-1} \qquad [11:112]$$

$$\tilde{\tilde{V}}_{0.1,1^2\ldots 1^{k-1}} = V_{0.1,1^2,1\ldots 1^{k-1}}\frac{N}{N-k} \qquad [11:113]$$

Accordingly the shrunken values of the squared correlations are

$${}_{s}r^2_{01} = {}_{s}r^2_{0\Delta 1} = \frac{(N-1)r^2_{0\Delta 1}-1}{N-2} \qquad [11:114]$$

Correction in r^2 for shrinkage,- linear regression, cf (12:36)

$${}_{s}r^2_{0\Delta 1,1^2} = \frac{(N-1)r^2_{0\Delta 1,1^2}-2}{N-3} \qquad [11:115]$$

Correction in r^2 for shrinkage,- quadric regression

$${}_{s}r^2_{0\Delta 1,1^2\ldots 1^j} = \frac{(N-1)r^2_{0\Delta 1,1^2\ldots 1^j}-j}{N-j-1} \qquad [11:116]$$

Correction in r^2 for shrinkage,- parabolic regression of jth degree, cf (12:36)

$$\epsilon^2_{01} = {}_{s}\eta^2_{01} = \frac{(N-1)\eta^2_{01}-k+1}{n-k} = 1-\frac{N-1}{N-k}(1-\eta^2_{01}) \qquad [11:117]$$

Correction in η^2 for shrinkage,- k classes (Kelley, 1935 and Peters and Van Voorhis, 1940, Ch. XI and XIII), also called correction for fine categories

If a series of r^2 values determined from linear, quadric, cubic, etc. regressions are computed that regression yielding an ${}_{s}r^2$ most nearly equal to ϵ^2 is the optimal regression and the ${}_{s}r^2$ cor-

responding is optimal. For ${}_{s}r^2 < \epsilon^2$ the regression employed does not get all that is of significance, and for ${}_{s}r^2 > \epsilon^2$ the regression used has presumably profited by chance.

We may also test any obtained ${}_{s}r^2$ by the variance ratio

$$F_{k-j-1,\,N-k} = \frac{(\eta_{01}^2 - r_{0\Delta 1,1^2\ldots 1^j}^2)/(k-j-1)}{(1-\eta_{01}^2)/(N-k)} \qquad [11:118]$$

To test adequacy of ${}_{s}r_{0\Delta 1,1^2\ldots 1^j}^2$

As here employed j equals the degree of the parabolic regression employed and k is the number of classes in the predictor variable.

For some purposes the variance error of ϵ^2, or of ϵ will suffice. As derived by Kelley (*op. cit.*) these are

$$V(\epsilon^2) = \frac{(1-\epsilon^2)^2}{N-1}\left[\frac{2(k-1)}{N-k} + 4\epsilon^2\right] \quad \ldots\ldots [11:119]$$

$$V_\epsilon = \frac{V(\epsilon^2)}{4\epsilon^2} \quad \text{(Valid only when } \epsilon^2 \text{ is not small)} \qquad [11:120]$$

Multiple and partial correlation ratios may be computed for multivariate categorical series. A three-variable illustration is given by Kelley in Rietz *Handbook* (1924).

CHAPTER XII

THE GENERAL MULTIPLE LINEAR REGRESSION PROBLEM

SECTION 1. A STATEMENT OF RELATIONSHIPS IN CONNECTION WITH STANDARD SCORE VARIABLES

Let it be desired to estimate X_0 from a knowledge of n independent measures, X_1, $X_2, \ldots, X_n$. The usual notation will be used for variances, standard deviations, correlations, covariances. Standard scores are defined as earlier and designated with the letter z: $z = (X-M)/\sigma$. The desired linear equation of estimate, or regression equation, is

$$\bar{X}_0 \equiv X_{0\Delta 12\ldots n} = a + b_1X_1 \quad b_2X_2 + \ldots + b_nX_n \qquad [12:01]$$

Multiple regression equation

When involving deviation from the mean scores this becomes

$$\bar{x}_0 \equiv x_{0\Delta 12\ldots n} = b_1x_1 + b_2x_2 + \ldots + b_nx_n \qquad [12:02]$$

and when employing standard scores it becomes

$$\bar{z} \equiv z_{0\Delta 12\ldots n} = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_n z_n \qquad [12:03]$$

Standard score form of regression equation

in which there is no constant term, for were there one the estimate of z_0 corresponding to the mean of the right-hand member, treated as a single variable, would not equal zero, which we know must be the case from the two-variable problem.

$$b_1 = \beta_1 \frac{\sigma_0}{\sigma_1}; \quad b_2 = \beta_2 \frac{\sigma_0}{\sigma_2}; \ldots; \quad b_n = \beta_n \frac{\sigma_0}{\sigma_n} \qquad [12{:}04]$$

$$a = M_0 - b_1 M_1 - b_2 M_2 - \ldots - b_n M_n \quad . \; . \; . \; . \; [12{:}05]$$

When [12:03] is employed the error of an estimate is $z_{0.12\ldots n}$,

$$z_{0.12\ldots n} = z_0 - z_{0\Delta 12\ldots n}$$

$$= z_0 - \beta_1 z_1 - \beta_2 z_2 - \ldots - \beta_n z_n \qquad [12{:}06]$$

The variance error of estimate, which is to be minimal, by the proper selection of the β's, is

$$k^2_{0.12\ldots n} = V(z_{0.12\ldots n}) = 1 + \beta_1^2 + \beta_2^2 + \ldots$$

$$+ \beta_n^2 - 2\beta_1 r_{01} - \ldots - 2\beta_n r_{0n} + 2\beta_1\beta_2 r_{12} + \ldots \; [12{:}07]$$

In the two-variable problem, involving z_0 and y, the y being a variable whose mean is zero and whose variance is V_y, we have

$$z_{0\Delta y} = r_{0y} \frac{1}{\sigma_y} y, \text{ and}$$

$$V(z_{0\Delta y}) = r^2_{0y} \frac{1}{V_y} V_y = r^2_{0y} \quad . \; . \; . \; . \; . \; . \; [12{:}08]$$

Accordingly, by treating the entire right-hand member of [12:03] as a single variable, we obtain

$$V(z_{0\Delta 12\ldots n}) = r^2_{0\Delta 12\ldots n} \qquad [12{:}09]$$

As with the two-variable problem,

$$z_0 = z_{0\Delta 12\ldots n} + z_{0.12\ldots n} \quad \ldots\ldots [12{:}10]$$

in which the two terms in the right-hand member are independent, so that, computing the variance of z_0 yields

$$1 = r^2_{0\Delta 12\ldots n} + k^2_{0.12\ldots n} \quad \ldots\ldots [12{:}11]$$

The error of estimate $(z_0 - z_{0\Delta 12\ldots n})$ is uncorrelated with z_1, for otherwise z_1, already used in the regression equation [12:03], could be employed again to reduce the error of estimate, but this is impossible because the β's in [12:03] are such as to yield a minimal variance error of estimate. We thus have, for the covariance between z_1 and $(z_0 - z_{0\Delta 12\ldots n})$

$$c[(z_0 - z_{0\Delta 12\ldots n})z_1] = c(z_0 z_1) - c(z_{0\Delta 12\ldots n} z_1) = 0$$

$$c(z_{0\Delta 12\ldots n} z_1) = c(z_0 z_1) = r_{01} \quad \ldots\ldots \quad [12{:}12]$$

Also

$$c(z_{0\Delta 12\ldots n} z_1) = c[z_1(\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_n z_n)]$$

$$= \beta_1 V(z_1) + \beta_2 c(z_1 z_2) + \ldots + \beta_n c(z_1 z_n)$$

or

$$r_{01} = \beta_1 + r_{12}\beta_2 + r_{13}\beta_3 + \ldots + r_{1n}\beta_n \qquad [12{:}13a]$$

Similarly

$$r_{02} = r_{12}\beta_1 + \beta_2 + r_{23}\beta_3 + \ldots + r_{2n}\beta_n \qquad [12{:}13b]$$

$$\begin{array}{ccccc} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{array}$$

$$r_{0n} = r_{1n}\beta_1 + r_{2n}\beta_2 + r_{3n}\beta_3 + \ldots + \beta_n \qquad [12{:}13n]$$

These n equations are the "condition equations" or "normal equations." Solved simultaneously, they yield the requisite values of the β's.

Thence substitution in [12:04] and [12:05] provides the constants of equation [12:01], which is the form needed for practical use.

$r^2_{0\Delta 12\ldots n}$ can be computed from $k^2_{0.12\ldots n}$, [12:11], which can be gotten from [12:07], or, very expeditiously, from [12:14] following:

We note that

$$c(z_0 z_{0\Delta 12\ldots n}) = c[z_0(\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_n z_n)]$$

$$= r_{01}\beta_1 + r_{02}\beta_2 + \ldots + r_{0n}\beta_n$$

but since, noting [12:08],

$$c(z_0 z_{0\Delta 12\ldots n}) = \sigma_{z_0}\sigma_{z_{0\Delta 12\ldots n}} r_{0.12\ldots n} = r^2_{0\Delta 12\ldots n}$$

we have

$$r^2_{0\Delta 12\ldots n} = r_{01}\beta_1 + r_{02}\beta_2 + \cdots + r_{0n}\beta_n \quad [12:14]$$

Many methods have been devised for solving the simultaneous equations [12:13]. Among them are the following, which have distinctive features of one sort or another: The determinantal solution, as given in Section 3 of this chapter, parallels and reveals theoretical development and is not uneconomical if four or fewer variables are involved. The Doolittle (1878) method, (see also Dwyer, 1941, *Doolittle* and, 1941, *Solution*), one of the many variations of which is given in the next section, (This modification was suggested in part by an article by Cowden, 1943) is generally serviceable, though it becomes rather laborious with 20 variables or more. Aitken's (1937) method of pivotal condensation has much similarity with and many of the merits of Doolittle procedure. The Kelley-Salisbury

(1926), (See also Kelley-McNemar, 1929, and Tolley-Ezekiel, 1927), or other iterative procedure, is designed for the many variable problem.

SECTION 2. A MODIFIED DOOLITTLE SOLUTION

We here give a four-variable problem solved by a modified Doolittle method, expressed in terms of symbols so that what has happened at every step is apparent. The multiple-regression problem involving more than four variables does not involve any principles not covered in the four-variable illustration.

A Doolittle solution based upon gross scores can be made. This would immediately yield the equation [12:01]. There are two disadvantages in this procedure: It involves one more variable, namely a, than a calculation based upon deviation scores, and it yields an answer [12:01] which holds but part of the information commonly desired, for a knowledge of means and variances is frequently pertinent to a problem in addition to knowledge of the regression equation. It is therefore judged to be generally economical to compute means, variances, and covariances in the first instance and then to perform a Doolittle solution with these measures, leading to equation [12:02]. Adding the constant a, [12:05] yields [12:01] which is generally the most serviceable form of the regression equation for practical use. It is readily established that the condition equations when deviation from mean scores are used differ slightly from those consequent to standard scores, [12:13], and are, for the four-variable problem, as follows:

(1) $b_1V_1 + b_2c_{12} + b_3c_{13} = c_{01}$ cf. Table XIII A row 1′

(2) $b_1c_{12} + b_2V_2 + b_3c_{23} = c_{02}$ cf. Table XIII A row 20

(3) $b_1c_{13} + b_2c_{23} + b_3V_3 = c_{03}$ cf. Table XII A row 30

In view of the fact that the coefficients of the unknown b's enter symmetrically (a situation leading to grammian determinants) the solution of these particular simultaneous equations becomes relatively simple. In the modified Doolittle layout of Table XII A the coefficients b_1, b_2, b_3, and the equality sign have been dropped and certain elements added (arising from relationships in matrix algebra connected with the identity matrix). The student can identify the steps of the Doolittle process by reference to steps (4) and (9) herewith. The Doolittle method is simply a convenient arrangement of steps for the solution of simultaneous equations in which the coefficients of the unknowns are symmetrical. Dividing the terms of equation (1) by $(-V_1)$ we obtain

(4) $-b_1 - b_2b_{21} - b_3b_{31} = -b_{01}$ cf. Table XII A row 1′

Multiply (1) by the second coefficient of (4), namely $-b_{21}$, obtaining

(5a) $-b_1V_1b_{21} - b_2c_{12}b_{21} - b_3c_{13}b_{21} = -c_{01}b_{21}$

(5b) $-b_1c_{12} - b_2r_{12}^2V_2 - b_3r_{13}r_{12}\sigma_3\sigma_2 = -r_{01}r_{12}\sigma_0\sigma_2$

which, using a notation in which Δ in a subscript means "dependent upon" is

(5) $-b_1c_{12} - b_2V_{2\Delta 1} - b_2c_{23\Delta 1} = -c_{02\Delta 1}$ cf. Table XII A row 21

Adding (2) and (5) yields an equation which is void of the b_1 term,

(6a) $b_2(V_2 - V_{2\Delta 1}) + b_3(c_{23} - c_{23\Delta 1}) = c_{02} - c_{02\Delta 1}$,

which, using a dot in a subscript to mean "independent of" may be written

(6) $b_2V_{2.1} + b_3c_{23.1} = c_{02.1}$ cf. Table XII A row 2

Dividing by $(-V_{2.1})$ we have,

(7) $-b_2 - b_3 b_{32.1} = -b_{02.1}$ cf. Table XII A row 2′

Multiplying (1) by the third coefficient of equation (4), namely $(-b_{31})$, we obtain,

(8) $-b_1 c_{13} - b_2 c_{23\Delta 1} - b_3 V_{3\Delta 1} = -c_{03\Delta 1}$ cf. Table XII A row 31

Adding (3) and (8) gives a second equation void of the b_1 term

(9) $b_2 c_{23.1} + b_3 V_{3.1} = c_{03.1}$ cf. Table XII A sum of rows 30 and 31

Equation (6) and (9) represent a three-variable problem in which the coefficients of the b's are symmetrical. Thus we can solve these in a similar manner, obtaining a final equation yielding b_3, which in the fuller notation of Table XII A is $b_{03.12}$.

Table XII A gives a Doolittle solution for a four-variable problem in algebraic notation, modified to give all the regression coefficients, thus obviating a "backward solution," and to give additional essential constants in rows B and C. When taking the square root of the squared partial correlation coefficients the signs of the corresponding regression coefficients in row 0 must be attached. When computing the multiple correlation from the relationship

$$r^2_{0\Delta 123} = 1 - k^2_{0.123} = 1 - \frac{V_{0.123}}{V_0} \quad \text{(See rows 0 and 00)} \quad [12:15]$$

the sign is always positive. As illustration of the dot and delta notation we note

$$V_0 = V_{0\Delta 1} + V_{0.1} \quad \text{(See [10:16])}$$

That is, the variance of x_0 may be set equal to the variance of a part which is dependent upon x_1 plus the variance of a part which is indepen-

dent of x_1. And again,

$$c_{01} = c_{01\Delta 2} + c_{01.2} \quad \ldots\ldots\ldots [12{:}16]$$

or the covariance c_{01} is equal to a part which is dependent upon x_2 plus a part which is independent of x_2. Since

$$c_{01.2} = (r_{01} - r_{02}r_{12})\sigma_0\sigma_1 \quad \ldots\ldots [12{:}17]$$

[12:16] may be written

$$r_{01}\sigma_0\sigma_1 = r_{02}r_{12}\,\sigma_0\sigma_1 + (r_{01} - r_{02}r_{12})\sigma_0\sigma_1,$$

so that

$$c_{01\Delta 2} = r_{02}r_{12}\,\sigma_0\sigma_1 \quad \ldots\ldots\ldots [12{:}18]$$

The addition of any number of the same secondary subscripts to each symbol does not change this relationship.

$$\frac{c_{01}}{V_1} = b_{01} \quad \ldots\ldots\ldots [12{:}19]$$

The addition of any number of the same secondary subscripts does not change this relationship.

$$b_{01} - b_{21}b_{02.1} = b_{01.2} \quad \ldots\ldots\ldots [12{:}20]$$

The addition of similar secondary subscripts to every symbol does not change this relationship.

The condition equations whose solution is provided for in Table XII A are numbers (1), (2), and (3). For future reference we add a symmetrical fourth row and write the coefficients of the unknowns in the accompanying augmented matrix:

TABLE XII A

SYMBOLIC EXPRESSION OF CONSTANTS DERIVED IN MODIFIED DOOLITTLE SOLUTION, 4 VARIABLES

	MULTIPLIERS	x_1	x_2	x_3	x_0	COLUMN I	COLUMN II	COLUMN III	CHECK COLUMN
$10 \equiv 1$		V_1	c_{12}	c_{13}	$-c_{01}$	-1			SUM
$1'$		-1	(21)	(31)	(01)	$1/V_1$			
20			V_2	c_{23}	$-c_{02}$	0	-1		SUM, INCLUDING c_{12}
21	$-b_{21}$		$-V_{2\Delta 1}$	$-c_{23\Delta 1}$	$c_{02\Delta 1}$	b_{21}	0		
2			$V_{2.1}$	$c_{23.1}$	$-c_{02.1}$	b_{21}	-1		CHECK CELL
$2'$			-1	(32)	(02)	$-b_{21}/V_{2.1}$	$1/V_{2.1}$		
30				V_3	$-c_{03}$	0	0	-1	SUM, INCLUDING c_{13}, c_{23}
31	$-b_{31}$			$-V_{3\Delta 1}$	$c_{03\Delta 1}$	b_{31}	0	0	
32	$-b_{32.1}$			$-V_{3.1\Delta 2.1}$	$c_{03.1\Delta 2.1}$	$-b_{21}b_{32.1}$	$b_{32.1}$	0	
3				$V_{3.12}$	$-c_{03.12}$	$b_{31.2}$	$b_{32.1}$	-1	CHECK CELL
$3'$				-1	(03)	$-b_{31.2}/V_{3.12}$	$-b_{32.1}/V_{3.12}$	$1/V_{3.12}$	

	MULTI-PLIERS	x_1	x_2	x_3	x_0	COLUMN I	COLUMN II	COLUMN III	CHECK COLUMN
00					V_0	0	0	0	SUM INCLUDING $-c_{01}, -c_{02}, -c_{03}$
01	b_{01}				$-V_{0\Delta 1}$	$-b_{01}$	0	0	
02	$b_{02.1}$				$-V_{0.1\Delta 2.1}$	$b_{21}b_{02.1}$	$-b_{02.1}$	0	
03	$b_{03.12}$				$-V_{0.12\Delta 3.12}$	$b_{31.2}b_{03.12}$	$b_{32.1}b_{03.12}$	$-b_{03.12}$	
0					$V_{0.123}$	$-b_{01.23}$	$-b_{02.13}$	$-b_{03.12}$	LAST CHECK CELL
0′					-1	$b_{01.23}/V_{0.123}$	$b_{02.13}/V_{0.123}$	$b_{03.12}/V_{0.123}$	
1×1′						$-1/V_1$			
2×2′						$-b_{21}^2/V_{2.1}$	$-1/V_{2.1}$		
3×3′						$-b_{31.2}^2/V_{3.12}$	$-b_{32.1}^2/V_{3.12}$	$-1/V_{3.12}$	
0×0′=A						$-b_{01.23}^2/V_{0.123}$	$-b_{02.13}^2/V_{0.123}$	$-b_{03.12}^2/V_{0.123}$	
SUM =B						$-1/V_{1.023}$	$-1/V_{2.013}$	$-1/V_{3.012}$	
C = A/B						$r_{01.23}^2$	$r_{02.13}^2$	$r_{03.12}^2$	

TABLE XII A
(CONTINUED)

$$\mathcal{B} = \begin{Vmatrix} V_1 & c_{12} & c_{13} & -c_{01} \\ c_{12} & V_2 & c_{23} & -c_{02} \\ c_{13} & c_{23} & V_3 & -c_{03} \\ -c_{01} & -c_{02} & -c_{03} & V_0 \end{Vmatrix} \quad \ldots\ldots\ldots [12{:}21]$$

In this arrangement the row and column connected with the criterion variable are placed last and the student will find it generally advantageous to arrange the predictor variables so that they are in an order, x_1, x_2, x_3, judged to be in decreasing order of importance as predictors. The advantage of this order is twofold. First, the crucial regression coefficient to test for significance is then $b_{03.12}$, which is the easiest to obtain. Precisely to test for the significance of $b_{03.12}$ one requires the few computations of column III. Those of columns I and II can be made later or not at all, as desired, depending upon the outcome of the test for significance of $b_{03.12}$. Second, should $b_{03.12}$ prove nonsignificant one would then desire the regression equation omitting x_3 and, it will be noted, most of the computational work for this is already at hand in the columns and rows of Table XII A that do not involve x_3.

In the cell at the intersection of row 1′ and column x_2 is the entry (21), which is to say that the numerical value found for this cell need not be recorded in this cell, it being in a more useful position if recorded in the multiplier column, row 21. Similarly for (31), (01), (32), etc.

Tables XII B and XII C provide the numerical data for the four-variable sample problem solved in Table XII E.

TABLE XII B

SAMPLE 4-VARIABLE MULTIPLE CORRELATION PROBLEM
Means and Variances,– N = 125

	X_1	X_2	X_3	X_0
M	52.20	43.60	45.40	68.15
V	92.80	149.92	203.08	110.32
σ	9.63	12.24	14.25	10.50

TABLE XII C

AUGMENTED VARIANCE AND COVARIANCE MATRIX
($-c_{0i}$ is entered)

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_0 \end{array} \left\| \begin{array}{cccc} x_1 & x_2 & x_3 & x_0 \\ 92.80 & 47.62 & 31.70 & -27.71 \\ & 149.92 & 10.47 & -28.92 \\ & & 203.08 & -20.05 \\ & & & 110.32 \end{array} \right\| = \mathfrak{B}$$

TABLE XII D

AUGMENTED CORRELATION MATRIX
($-r_{0i}$ is entered)

$$\left\| \begin{array}{cccc} 1.000 & .404 & .231 & -.274 \\ & 1.000 & .060 & -.225 \\ & & 1.000 & -.134 \\ & & & 1.000 \end{array} \right\| = \mathfrak{a}$$

TABLE XII E

MODIFIED DOOLITTLE NUMERICAL SOLUTION,--4 VARIABLES

	MULTI-PLIERS	x_1	x_2	x_3	x_0	COLUMN I	COLUMN II	COLUMN III	CHECK COLUMN
10=1		92.8000	47.6200	31.70000	−27.7100	−1.000000			144.4100
1′		−1.	(21)	(31)	(01)	.010776			xxx
20			149.9200	10.4700	−28.92	.0	−1.000000		179.0900
21	−.513153		−24.4363	−16.2670	14.2195	.513153	.0		−74.1044
2			125.4837	−5.7970	−14.7005	.513153	−1.000000		104.9856
2′									104.9857
2′			−1.	(32)	(02)	−.004089	.007969		xxx
30				203.0800	−20.0500	.0	.0	−1.000000	225.2000
31	−.341595			−10.8286	9.4656	.341595	.0	.0	−49.3297
32	.046197			−.2678	−.6791	.023706	−.046197	.0	4.8500
3				191.9836 $V_{3.12}$	−11.2635	.365301	.046197	−1.000000	180.7203 180.7201
3′				−1.	(03)	−.001903	.000241	.005209	xxx

	MULTI-PLIERS	x_1	x_2	x_3	x_0	COLUMN I	COLUMN II	COLUMN III	CHECK COLUMN
00					110.3200	.0	.0	.0	33.6400
01	.298603				−8.2743	−.298603	.0	.0	−43.1213
02	.117148				−1.7221	.060115	−.117148	.0	12.2989
03	.058669				−.6608	.021432	−.002710	−.058669	10.6027
0					99.6628 $\{V_{0.123}$	$\{-$.217056 $\{-b_{01.23}$	$\{-$.119858 $\{-b_{02.13}$	$\{-$.058669 $\{-b_{03.12}$	99.6629 99.6628
0′					−1.	.002178	.001203	.000589	
1x1′						−.010776			
2x2						−.002098	−.007969		
3x3						−.000695	−.000011	−.005209	
0x0′=A						−.000473	−.000144	−.000035	
SUM=B						−.014042 $\{-1/V_{1.023}$	−.008142 $\{-1/V_{2.013}$	.005244 $\{-1/V_{3.012}$	
A/B=C						.033685 $\{r^2_{01.23}$	.017725 $\{r^2_{02.13}$	.006674 $\{r^2_{03.12}$	

TABLE XII E
(CONTINUED)

It is to be noted that the check column applies to the work between the double vertical rules. To obtain a check of the b's it is necessary to substitute them in one of the condition equations and see if it reduces to zero. This is most readily done employing the constants of the first condition equation, as recorded in row 10 of Table XII E. In symbols the equation is

$$b_{01.23}V_1 + b_{02.13}c_{12} + b_{03.12}c_{13} - c_{01} = 0$$

Thus we have

$$.217055 \times 92.80 + .119858 \times 47.62 + .058671 \times 31.70 - 27.71 = .0002$$

This being zero within the accuracy of the computational procedure constitutes a check.

There is no independent check upon the computational work below row O, so it is important that this be performed with extra caution.

The multiple correlation squared, as given by [12:15], is .0966, but this has a bias and should be corrected for shrinkage by [12:36]. The corrected value is .0742, or ${}_{s}r_{0\Delta 123} = .2724$. Since the single correlation r_{01} (see Table XII C) equals .274, it is futile to use all three predictors in this particular problem, but we shall proceed to apply precise tests to the multiple correlation coefficient and to the regression coefficients. Utilizing [12:37], we have

$$F_{3,121} = \frac{.0966 \times 121}{.9034 \times 3} = 4.3128$$

This being greatly in excess of 1.0, we may be sure that $r^2_{0\Delta 123}$ is not a chance deviation from zero. The computation of P for this variance ratio will only confirm this belief. We do, in fact, find that $P = .038$.

To test the significance of a multiple-regression coefficient we have a variance ratio of the sort [10:78] by adding secondary subscripts and noting that the error variance has $(N-n-1)$ degrees of freedom.

$$F_{1(N-n-1)} = \frac{(b_{0i.12\ldots)i(\ldots n} - \tilde{b})^2 V_{i.12\ldots)i(\ldots n} (N-n-1)}{V_{0.12\ldots n}} \quad [12:22]$$

See also [12:24]

For the two-variable problem we have

$$V_i = V_{i\Delta 0} + V_{i.0} = V_i r_{0i}^2 + V_{i.0}$$

$$V_i = \frac{V_{i.0}}{1-r_{0i}^2}$$

Accordingly, adding secondary subscripts $12\ldots)i(\ldots n$ throughout, we obtain

$$V_{i.12\ldots)i(\ldots n} = \frac{V_{i.012\ldots)i(\ldots n}}{1-r_{0i.12\ldots)i(\ldots n}^2} \quad [12:23]$$

Substituting in [12:22] yields

$$F_{1(N-n-1)} = \frac{(b_{0i.12\ldots)i(\ldots n} - \tilde{b})^2 V_{i.012\ldots)i(\ldots n} (N-n-1)}{V_{0.12\ldots n} (1-r_{0i.12\ldots)i(\ldots n}^2)} \quad [12:24]$$

Since the negative of the reciprocal of the numerator variance is given in row B and the square of the partial correlation coefficient in row C, all the necessary constants are available to apply this test. Also, since this is a variance ratio having one d.o.f. in the numerator, its square root is a critical ratio and, if the d.o.f. of the denominator is not small, then the square

root of this F may be treated as a deviate in a unit normal distribution.

The variance ratio to test if $b_{03.12}$ is a chance deviation from zero is

$$F_{1,121} = \frac{(.058671)^2 \times 121}{.005244 \times 99.6628 \times .993326} = .802215$$

$$\sqrt{F_{1,121}} = .8957 \text{ and } P = .370.$$

Similarly for $b_{02.13}$ we obtain $P = .139$, and for $b_{01.23}$ we obtain $P = .040$.

It is equally effective to test the partial correlation coefficient as a chance deviation from zero as to test the partial regression coefficient as a chance deviation from zero. We have $r_{03.12} = \sqrt{.006674} = .0817$. Using the Fisher r-into-z transformation, we obtain $z_{03.12} = .08188$. This has a standard error of .09129 $(= 1/\sqrt{121-1})$. The critical ratio is .8969 and to three decimal places $P = .370$, which is equal to that previously obtained for $b_{03.12}$ as a chance deviation from zero. These two tests are alternatives. The r-into-z procedure gives a P of .142 for $r_{02.13}$ which is to be compared with the previous value .139. Similarly for $r_{01.23}$ we obtain $P = .042$, to be compared with .040.

For the significance of $(M_0-\tilde{M}_0)$ we have a test similar to [10:38]. We note that $(n+1)$ linear restrictions have been imposed upon the X_0 measures. They are the restrictions inherent in M_0, c_{01}, $c_{02}, \ldots, c_{0n}$. Thus we have

$$F_{1(N-n-1)} = \frac{(M_0 - \tilde{M}_0)^2 \; (N-n-1)}{V_{0.12...n}} \qquad \begin{array}{l}\text{Variance ratio}\\ \text{test of } M_0\\ \text{in the case}\\ \text{of multiply}\\ \text{correlated data}\end{array} \quad [12:25]$$

SECTION 3. THE DETERMINANTAL EXPRESSION OF THE SOLUTION OF NORMAL EQUATIONS

In connection with the condition equations [12:13] let us write down the augmented correlation major determinant A, or Δ. This corresponds to [12:21] except that the variables here are standard scores.

$$A = \Delta = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1n} & r_{01} \\ r_{21} & 1 & \cdots & r_{2n} & r_{02} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 & r_{0n} \\ r_{01} & r_{02} & \cdots & r_{0n} & 1 \end{vmatrix} \quad \ldots \quad [12:26]$$

The following relationships hold, in which, of course, Δ_{0i} is the determinant obtained by crossing out the last row and the variable i column of the major determinant A, and $A_{0i} = (-1)^{n+i+1}\Delta_{0i}$.

$$k^2_{0.12...n} = \frac{\Delta}{\Delta_{00}} \quad . \; . \; . \; . \; . \; . \; . \; [12:27]$$

$$r^2_{0\Delta12...n} = 1 - \frac{\Delta}{\Delta_{00}} \quad . \; . \; . \; . \; . \; [12:28]$$

$$\beta_1 = \frac{-A_{01}}{A_{00}} = \frac{-\Delta_{01}}{\Delta_{00}}; \quad \beta_2 = \frac{-A_{02}}{A_{00}} = \frac{\Delta_{02}}{\Delta_{00}}; \quad \beta_3 = \frac{-A_{03}}{A_{00}} = \frac{-\Delta_{03}}{\Delta_{00}}; \text{ etc.}$$

$$\beta_i = \frac{-A_{0i}}{A_{00}} = (-1)^{n+i+1}\frac{\Delta_{0i}}{\Delta_{00}} \qquad [12:29]$$

$$r^2_{0i.12\ldots)i(\ldots n} = \beta_{0i.12\ldots)i(\ldots n}\,\beta_{i0.12\ldots)i(\ldots n} \qquad [12:30]$$

$\beta_{0i.12\ldots)i(\ldots n}$ is the expanded notation for β_i, as given by [12:29]. The reversed parentheses, $)i($, indicate the i is omitted in the series of secondary subscripts $1,2,\ldots,n$. Since $A_{0i} = A_{i0}$, and the sign factor for the position $(0,i)$ is the same as for the position $(i,0)$, we can write

$$r^2_{0i.12\ldots)i(\ldots n} = \frac{A^2_{0i}}{A_{00}A_{ii}} = \frac{\Delta^2_{0i}}{\Delta_{00}\Delta_{ii}} \quad \text{The squared partial correlation coefficient} \qquad [12:31]$$

When extracting the square root to obtain

$$r_{0i.12\ldots)i(\ldots n}$$

the sign is the sign of $-A_{0i}$, or in other words it is the sign of β_i.

Since β_i vanishes with this partial correlation, the significance of β_i as a deviation from zero is exactly the significance of

$$r_{0i.12\ldots)i(\ldots n}$$

as a deviation from zero. The distribution of a partial correlation coefficient is the same as that of any other correlation coefficient, when proper allowance is made for its reduced degrees of freedom. The total correlation coefficient, r_{01}, has $(N-2)$ degrees of freedom; its variance error is $(1-r^2_{01})^2/(N-2)$, as given by [10:46], of Chapter X; and when transformed into a Fisher z, see [10:43], it has a variance error, see [10:44], of $1/(N-3)$. Accordingly, in the case of this partial correlation in this $(n+1)$ variable problem,

$$V(r_{0i.12\ldots)i(\ldots n}) = \frac{(1-r^2_{0i.12\ldots)i(\ldots n})^2}{N-n-1} \qquad [12:32]$$

Variance error of partial r

$$V(z \text{ from partial } r) = \frac{1}{N-n-2} \qquad [12:33]$$

Variance error of partial z

Formula [12:32] gives the variance error of the correlation between X_0 and X_1 for fixed values of X_2, X_3,...X_n. This is to say that the sample in question is one of an infinite number, all of which have the identical X_2, X_3,...,X_n values and in the identical pairings, X_2 with X_3, X_2 with X_4, etc., as in this observed sample. Thus the only variables which change from sample to sample are X_0 and X_1. The partial correlation $r_{01.23\ldots n}$, read "the correlation between X_0 and X_1 for fixed values of variables X_2, X_3, ...,X_n," has no meaning unless X_2, X_3,...,X_n are fixed. This partial correlation is indeterminate unless X_2, X_3,...,X_n have fixed values, so no formula based on the correlations in a single sample, for the variance error of a partial correlation coefficient (or of a regression coefficient in a multiple variable problem) is possible under totally free sampling.

Should we have a number of totally free samplings and compute an $r_{01.23\ldots n}$ for each, these partial r's would vary in meaning and would vary in magnitude and their variance error could be computed, but, so far as the writer is aware, this has never been done. We must therefore look upon the partial correlation coefficient (and the multiple regression coefficient) as having a known variance error only when all variables, except the two primary ones in question, have fixed values.

The variance of b_1, as given by [12:34] is

its variance when all variables except X_0 and X_i have the fixed values of the observed sample, thus by parity with [11:79],

$$V(b_i) = V(b_{0i.12\ldots)i(\ldots n}) = \frac{V_{0.12\ldots n}}{V_{i.12\ldots)i(\ldots n}(N-n-1)}$$

$$= \frac{V_0 k^2_{0.12\ldots n}}{V_i k^2_{i.12\ldots)i(\ldots n}(N-n-1)} \qquad [12:34]$$

Variance error of a multiple regression coefficient

The divergence of the observed value, b_i, from any theoretical value, $\tilde{b}_i$, is to be tested by the variance ratio

$$F_{1(N-n-1)} = \frac{(b_i - \tilde{b}_i)^2 \; V_i k^2_{i.12\ldots)i(\ldots n}(N-n-1)}{V_0 k^2_{0.12\ldots n}} \qquad [12:35]$$

(See also [12:24])

When, as is common in multiple regression problems, $(N-n-1)$ is large, $\sqrt{F_{1(N-n-1)}}$ is sensibly the same as the x deviate in a unit normal distribution, so that double the area to the right of x is a close approximation to the variance ratio P.

The multiple correlation coefficient is a biased measure, being larger then is to be anticipated if the obtained regression is applied to a new sample and the correlation with a similar criterion computed. The amount of this bias is known, (see Fisher, 1923, and Wherry, 1939), and the following formula applies, in which ${}_sr^2$ is r^2 corrected for shrinkage:

$${}_sr^2_{0\Delta 12\ldots n} = \frac{(N-1)r^2_{0\Delta 12\ldots n} - n}{N-n-1} \qquad [12:36]$$

Giving correction in r^2 for shrinkage in the (n+1) variable problem, cf. [11:116]

The limiting values of multiple r are 0 and 1, so the form of distribution is not identical with that of a total correlation coefficient, so we shall not use the r-into-z technique, of Chapter X. We may compute a variance ratio, similar to [10:42], to test the hypothesis that true multiple r^2 is equal to 0. Corresponding to the variance equation,

$$V_0 = V_{0\Delta 12\ldots n} + V_{0.12\ldots n} = V_0 r^2_{0\Delta 12\ldots n} + V_0(1 - r^2_{0\Delta 12\ldots n})$$

is the degree of freedom equation,

$$(N-1) = n + (N-n-1)$$

The multiple correlation coefficient has as many degrees of freedom as independent variables, n, and the error variance has $N-n-1$, so that,

$$F_{n(N-n-1)} = \frac{r^2_{0\Delta 12\ldots n}(N-n-1)}{(1-r^2_{0\Delta 12\ldots n})(n)} \quad F \text{ to test hypothesis that } r_{0\Delta 12\ldots n} = 0 \qquad [12:37]$$

In addition to the relationships of this Section, further relationships are neatly expressed by means of matrices, as explained in Section 4.

SECTION 4. THE USE OF MATRICES IN EXPRESSING MULTIPLE CORRELATION RELATIONSHIPS

The treatment herewith assumes the elementary familiarity with matrices and determinants represented by the discussion of Chapter XIV, Section 2. The illustration herewith is based upon a four-variable problem, but the relationships given hold for a problem with any number of variables. Were there a simple way to evaluate determinants of fairly high order, the solution of regression equations by means of matrices and

determinants would undoubtedly be the standard method, for it brings into clear focus the underlying algebraic and geometric relationships. We start with an augmented matrix of type [12:38]. For condition equations [12:13] this is

$$\mathfrak{A} = \begin{Vmatrix} 1 & r_{12} & r_{13} & r_{01} \\ r_{12} & 1 & r_{23} & r_{02} \\ r_{13} & r_{23} & 1 & r_{03} \\ r_{01} & r_{02} & r_{03} & 1 \end{Vmatrix} \qquad \text{Augmented matrix} \quad [12:38]$$

The transpose of $\mathfrak{A}$ is $\mathfrak{A}'$ and its inverse is $\mathfrak{A}^{-1}$. We will designate the elements entering into A as a_{ij}. The cofactor of a_{ij} is A_{ij}, in which i refers to the row and j to the column. When $\mathfrak{A}$, $\mathfrak{A}'$, and $\mathfrak{A}^{-1}$ are treated as determinants they are designated A, A', and A^{-1}. The elements of $\mathfrak{A}^{-1}$ (or of A^{-1}) are a^{ji} (at the intersection of the j row and the i column).

$$a^{ji} = \frac{A_{ij}}{A} \qquad \text{Elements of the inverse matrix } \mathfrak{A}^{-1} \quad [12:39]$$

When the matrix is of predictor variables only we use a notation similar to the preceding, substituting R for A. Thus,

$$\mathfrak{R} = \begin{Vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{Vmatrix} \qquad \text{Predictor matrix} \quad [12:40]$$

The meanings of $\mathfrak{R}'$, $\mathfrak{R}^{-1}$, R, R', R^{-1}, r_{ij}, the element in $\mathfrak{R}$ (or R), R_{ij} its cofactor, r^{ji}, the element in $\mathfrak{R}^{-1}$ (or R^{-1}), will be obvious.

$$r^{ji} = \frac{R_{ij}}{R}$$ Elements of the inverse matrix $\mathcal{R}^{-1}$ [12:41]

The vector matrix of regression coefficients is β,

$$\beta = \begin{Vmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{Vmatrix} \equiv \begin{Vmatrix} \beta_{01.23} \\ \beta_{02.13} \\ \beta_{03.12} \end{Vmatrix} \quad . \; . \; . \; . \; . \; . \; . \; . \quad [12:42]$$

The vector matrix of correlations with the predictand is $\mathcal{r}$,

$$\mathcal{r} = \begin{Vmatrix} r_{01} \\ r_{02} \\ r_{03} \end{Vmatrix} \text{ and transpose } \mathcal{r}' = \begin{Vmatrix} r_{01} & r_{02} & r_{03} \end{Vmatrix} \quad [12:43]$$

The following relationships, which hold for any number of variables, will not be proven, but merely illustrated with the four-variable data of Section 2.

$$\beta = \mathcal{R}^{-1} \mathcal{r}$$ Yielding the matrix of β regression coefficients [12:44]

$$1 - r^2_{0\Delta 123} = k^2_{0.123} = \frac{A}{R} \quad . \; . \; . \; . \; . \; . \quad [12:45]$$

$$r^2_{0\Delta 123} = \beta \mathcal{r}' = \beta' \mathcal{r} \quad . \; . \; . \; . \; . \; . \; . \quad [12:46]$$

$$V_{\beta_i} = \frac{r^{ii} \; k^2_{0.123}}{N-4}$$ Variance error of a β multiple regression coefficient, -4 variables [12:47]

For the general case we have

$$V_{\beta_i} = \frac{r^{ii} k^2_{0.12\ldots n}}{N-n-1} \quad \text{Variance error of a } \beta \text{ multiple regression coefficient, -(n+1) variables (See also [12:34])} \qquad [12:48]$$

$$= \frac{A \quad R_{ii}}{R^2(N-n-1)} \quad \ldots\ldots\ldots\ldots \qquad [12:48a]$$

A derivation of this follows directly from the $\sigma^2_{b_1}$ formula on page 600 of Fisher (1922).

We also have

$$V_{b_i} = V_0 \; V_{\beta_i} \quad \text{Variance error of a multiple regression coefficient (see also [12:34])} \qquad [12:49]$$

The significance of the difference between two of the β coefficients entering into a single regression equation is sometimes desired. These are not independent, so a covariance term is present. The general equation is:

$$V(\beta_i - \beta_j) = \frac{(r^{ii} + r^{jj} - 2r^{ij})\; k^2_{0.12\ldots n}}{N - n - 1} \qquad [12:50]$$

Variance error of difference between β regression coefficients, -(n+1) variables

Further

$$F_{1,N-n-1} = \frac{(\beta_i - \beta_j)^2}{V(\beta_i - \beta_j)} \quad \text{Variance ratio to test the significance of } (\beta_i - \beta_j) \qquad [12:51]$$

Inherent in this equation is the relationship

$$r_{\beta_i \beta_j} = \frac{r^{ij}}{\sqrt{r^{ii} \quad r^{jj}}} \quad \ldots\ldots\ldots\ldots \qquad [12:52]$$

Since the criterion variable does not enter into the right-hand member we note that the correlation between regression coefficients is in-

dependent of the criterion, or predictand.

Under conditions of sampling such that V_i and V_j do not vary from sample to sample, we have

$$r_{b_i b_j} = r_{\beta_i \beta_j} \qquad \text{Correlation between regression coefficients} \qquad [12:53]$$

The data of the four-variable problem used in the modified Doolittle solution is now employed to illustrate the solution by matrices and determinants.

Using the values given in the augmented correlation matrix A, Table XII D, we evaluate as a determinant and find that $A = .714545$.

The Predictor Matrix $\mathfrak{R}$ is

$$\begin{array}{cccc} & (1) & (2) & (3) \\ \mathfrak{R} = & \left\| \begin{matrix} 1.000 & .404 & .231 \\ .404 & 1.000 & .060 \\ .231 & .060 & 1.000 \end{matrix} \right\| & & \end{array} \qquad \text{In which the elements are designated } r_{ij}$$

Evaluating as a determinant we obtain $R = .791022$. Utilizing [12:28] we obtain $r^2_{0\Delta 123} = .0967$, to be compared to the value .0966 obtained in Section 2.

The cofactors of the elements in $\mathfrak{R}$ are R_{ij}, as recorded in the adjoint matrix herewith:

$$\left\| \begin{matrix} .996400 & -.390140 & -.206760 \\ -.390140 & .946639 & .033324 \\ -.206760 & .033324 & .836784 \end{matrix} \right\|$$

Dividing each of these by R we obtain the elements of r^{ji} of the inverse matrix R^{-1} herewith:

$$\mathcal{R}^{-1} = \begin{Vmatrix} 1.259636 & -.493210 & -.261383 \\ -.493210 & 1.196729 & .042128 \\ -.261383 & .042128 & 1.057852 \end{Vmatrix}$$

Matrix multiplication yields the identity matrix $\mathcal{I}$, which is useful as a check upon the arithmetic accuracy of the work.

$$\mathcal{R}\,\mathcal{R}^{-1} = \begin{Vmatrix} 1.000000 & .000000 & .000001 \\ .000000 & 1.000000 & .000000 \\ .000001 & .000000 & 1.000000 \end{Vmatrix} = \mathcal{I}$$

The matrix of correlations with the criterion, x_0, is $\mathcal{r}$.

$$\mathcal{r} = \begin{Vmatrix} r_{01} \\ r_{02} \\ r_{03} \end{Vmatrix} = \begin{Vmatrix} .274 \\ .225 \\ .134 \end{Vmatrix}$$

The β regression coefficients matrix is

$$\beta = \mathcal{R}^{-1}\mathcal{r} = \begin{Vmatrix} .199143 \\ .193770 \\ .079612 \end{Vmatrix} = \begin{Vmatrix} \beta_{01.23} \\ \beta_{02.13} \\ \beta_{03.12} \end{Vmatrix} = \begin{Vmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{Vmatrix}$$

Since $\beta_i = b_i \dfrac{\sigma_0}{\sigma_i}$ we utilize the σ values of Table XII B and obtain

$$b_1 = \beta_1 \frac{\sigma_0}{\sigma_1} = .217134; \quad b_2 = \beta_2 \frac{\sigma_0}{\sigma_2} = .119901;$$

$$b_3 = \beta_3 \frac{\sigma_0}{\sigma_1} = .058661$$

which check with the values obtained from the modified Doolittle solution, for three and four-figure σ values were employed.

The r^{ji} values in R^{-1} provide us with the necessary information for testing the significance of the difference between two β's, formula [12:51]. For example:

$$(\beta_1 - \beta_2) = .0593$$

and

$$k^2_{0.123} = \frac{.714545}{.791022} = .903319$$

so that

$$V(\beta_1 - \beta_2) = \frac{[1.259636 + 1.196729 - 2(-.493210)](.903319)}{125 - 4}$$

$$= .0257$$

or σ.

$$\sigma_{(\beta_1 - \beta_2)} = .16.$$

Accordingly, the difference .0593 is not trustworthy.

CHAPTER XIII

SUNDRY STATISTICAL ISSUES AND PROCEDURES

SECTION 1. EVIDENCE OF PERIODICITY IN SHORT TIME SERIES (An abridgment of Kelley, 1943)

Many time series are of short duration because the recording of the event has, of necessity, proceeded for but a short time. Though it may be necessary that the student bide his time and wait for the years to pass in order to secure sufficient data, we should come to this conclusion only after having exhausted the possibilities of small sample theory. A method which applies to short series and which enables a judgment of the hypothesis with fiducial limits is important in that, whatever the time span, the issue can be investigated and the hypothesis substantiated or rejected within the fiducial limits set by the investigator.

Let the time variable be X_1 and let X_0 be the variable which we postulate to be a periodic function of X_1. Before investigating periodicity we should first take out any linear, quadric, or other non-periodic trend that may exist. We pause to note that the inclusion of "quadric, or other non-periodic" is a broadening, but we be-

lieve justifiable estension, of the usual definition of trend as that allowance which has to be made for the time such that the residuals shall be uncorrelated with the time.

For the purposes of precise test it is advantageous to have the trend represented by a line that places linear restrictions only upon X_0, as do the parabolic regression constants of Chapter XI, Section 4. Consider the following series of estimates of X_0:

$$\text{Estimate of } X_0 = M_0 \ldots X_1^0 \qquad \text{Regression} \qquad [13:01]$$

$$X_{0\Delta 1} = a + b\,X_1 \ldots \qquad \text{Linear regression} \qquad [13:02]$$

$$X_{0\Delta 1,1^2} = A + B\,X_1 + C\,X_1^2 \qquad \text{Quadric regression} \qquad [13:03]$$

$$X_{0\Delta 1,1^2,1^3} = \alpha + \beta\,X_1 + \gamma X_1^2 + \delta X_1^2 \qquad \text{Cubic regression} \qquad [13:04]$$

We test the need of [13:02] rather than [13:01], that is, we test the hypothesis $b = 0$, as in Chapter XI, Section 4. Or again we test the need of [13:03] rather than [13:02], that is the hypothesis $C = 0$, etc. When no trend is removed [13:01] provides the estimate of X_0 and the quantities related to X_1 in which we seek periodicity are $(X_0 - M_0)$. These have N-1 degrees of freedom. If a linear trend is removed by [13:02] the residual quantities, having N-2 degrees of freedom, in which we seek periodicity are

$$X_{0.1} \; (= X_0 - X_{0\Delta 1})$$

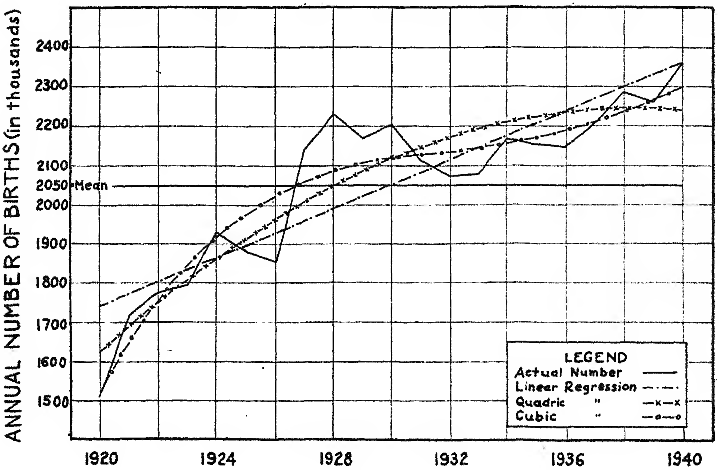
When a quadric trend is removed by [13:03] the residuals are $X_{0.1,1^2}$ and they have N-3 degrees of freedom.

A similar procedure, equation [13:04], could remove a cubic trend, but should be used with caution as a cubic is itself a close approxima-

tion to a complete cycle of a certain sine curve. The cubic curve of Chart XIII I, which cuts the quadric much as would a sine curve with a period of 16 years, is an illustration of this.

CHART XIII I

BIRTHS IN UNITED STATES 1920-1940



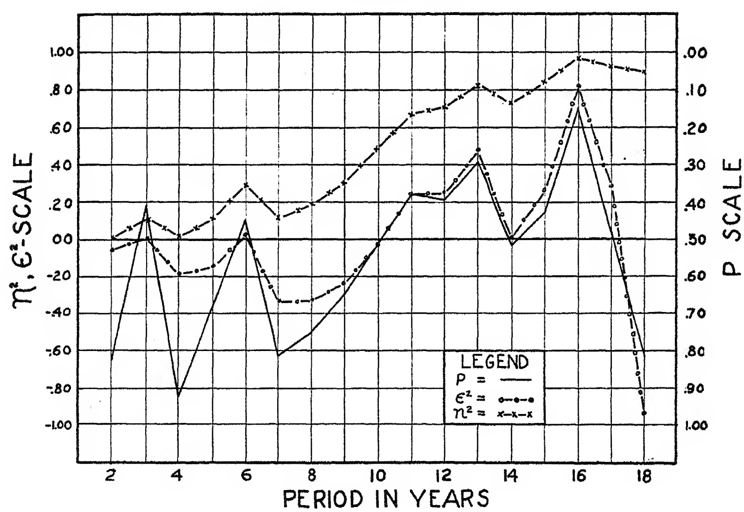
Using the data in the first two columns of Table XIII A, we find a quadric trend, (C of [13:03] $\neq 0$), and remove the same obtaining residuals $X_{0.1,1^2}$, which may be investigated for periodicity. We employ a periodogram technique with two modifications. The usual periodogram is as shown in the dot line of Chart XIII II. The raw correlation ratio squared, η^2, is plotted for successively greater periods, T. The argument is that in the neighborhood of the period yielding the largest η^2 will be found the actual period of the data. For this argument to be precise it is necessary that a probability statement be attached to the outcome. Thus in addition to, or instead of, the η^2 ordinate in the

periodogram we employ the ordinate $(1-P)$, the P being that from a variance ratio, the numerator variance being that of the means of the classes into which the data have been grouped when items of the same phase for the period T constitute a group, and the denominator variance is the customary residual, or error, variance.

A briefer process, yielding essentially the same picture, as illustrated in Chart XIII II, is to use as the ordinate the unbiased correlation ratio squared, ϵ^2. There lies in ϵ the additional value that it gives a quantitative statement of the correlation between time, treated as a periodical variable of period T, and the magnitudes, $X_{0.1,1^2}$, being investigated.

CHART XIII II

PERIODOGRAM



We start with N equally time-spaced observations X_0. We find that a quadric trend exists by the method of Chapter XI, Section 4, and remove it, giving N residuals, $X_{0.1,1^2}$, having $N-3$ degrees of freedom.

TABLE XIII A

BIRTHS IN THE UNITED STATES 1920-1940
(World Almanac, 1943, p. 472)

X_1 YEAR	X_0 BIRTHS	$\frac{X_{0\Delta 1,1^2}}{100}$	$\frac{X_{0.1,1^2}}{100}$	PHASES FOR PERIODS T					
				2	3	4	5	6	7
1920	1508874	16219	−1130	a	a	a	a	a	a
1	1714261	16885	258	b	b	b	b	b	b
2	1774911	17514	235	a	c	c	c	c	c
3	1792646	18105	− 179	b	a	d	d	d	d
4	1930614	18658	648	a	b	a	e	e	e
5	1878880	19174	− 385	b	c	b	a	f	f
6	1856068	19652	−1091	a	a	c	b	a	g
7	2137836	20093	1285	b	b	d	c	b	a
8	2233149	20496	1835	a	c	a	d	c	b
9	2169920	20862	837	b	a	b	e	d	c
1930	2203958	21190	850	a	b	c	a	e	d
1	2112760	21480	− 352	b	c	d	b	f	e
2	2074042	21733	− 993	a	a	a	c	a	f
3	2081232	21949	−1137	b	b	b	d	b	g
4	2167636	22127	− 451	a	c	c	e	c	a
5	2155105	22267	− 716	b	a	d	a	d	b
6	2144790	22370	− 922	a	b	a	b	e	c
7	2203337	22435	− 402	b	c	b	c	f	d
8	2286962	22462	408	a	a	c	d	a	e
9	2265558	22453	203	b	b	d	e	b	f
1940	2360399	22405	1199	a	c	a	a	c	g
			η^2 =	.004	.115	.014	.112	.298	.109
			$F_{(k-1)(N-2-k)}$ =	.073	1.042	.414	.439	1.104	.244
			P =	.825	.408	.923	.682	.449	.817
			ϵ^2 =	−.054	.005	−.183	−.142	.028	−.337

TABLE XIII A

BIRTHS IN THE UNITED STATES 1920-1940

(World Almanac, 1944, p. 472)

PHASES FOR PERIODS T										
8	9	10	11	12	13	14	15	16	17	18
a	a	a	a	a	a	a	a	a	a	a
b	b	b	b	b	b	b	b	b	b	b
c	c	c	c	c	c	c	c	c	c	c
d	d	d	d	d	d	d	d	d	d	d
e	e	e	e	e	e	e	e	e	e	e
f	f	f	f	f	f	f	f	f	f	f
g	g	g	g	g	g	g	g	g	g	g
h	h	h	h	h	h	h	h	h	h	h
a	i	i	i	i	i	i	i	i	i	i
b	a	j	j	j	j	j	j	j	j	j
c	b	a	k	k	k	k	k	k	k	k
d	c	b	a	l	l	l	l	l	l	l
e	d	c	b	a	m	m	m	m	m	m
f	e	d	c	b	a	n	n	n	n	n
g	f	e	d	c	b	a	o	o	o	o
h	g	f	e	d	c	b	a	p	p	p
a	h	g	f	e	d	c	b	a	q	q
b	i	h	g	f	e	d	c	b	a	r
c	a	i	h	g	f	e	d	c	b	a
d	b	j	i	h	g	f	e	d	c	b
e	c	a	j	i	h	g	f	e	d	c
.185	.312	.489	.665	.708	.826	.725	.837	.969	.920	.893
.360	.568	.958	1.590	1.541	2.368	1.016	1.468	6.240	1.130	.490
.750	.649	.512	.378	.391	.291	.519	.429	.152	.485	.814
-.331	-.238	-.022	.248	.248	.477	.012	.267	.314	.283	-.927

Let the period under investigation be T of the equally spaced time intervals. Of necessity $1 < T < (N-2)$ and the practical limits of utility of T, fractional or integral, may be taken such that $2 \leq T \leq (N-2)$, while if a zero trend, equation [13:01], is involved $2 \leq T \leq (N-1)$.

If N is exactly divisible by T, there are T classes each having N/T measures, $X_{0.1,1^2}$, of the same phase in it. If N is not exactly divisible by T, we have a variable number of measures from phase to phase and the number of classes will be some number which we call k, greater than N/T. Let the means of the $X_{0.1,1^2}$ measures in these classes be M_{0a}, $M_{0b}, \ldots M_{0k}$. The variance of these means, appropriately weighted with the number of measures in each class, is designated $V(M_{0k})$. The mean for each class is independent of that for every other class except that the mean of the means is equal to zero, so that there are just $(k-1)$ degrees of freedom. Letting i take all values from a to k, we can express the variable $X_{0.1,1^2}$ as equal to the sum of two independent parts. We have the following relationships:

$$X_{0.1,1^2} = M_{0i} + X_{0.1,1^2,i} \quad \ldots\ldots \quad [13:05]$$

$$N-3 = (k-1) + (N-2-k) \quad \text{D.O.F. equation} \quad [13:06]$$

$$V_{0.1,1^2} = V(M_{0K}) + V_{0.1,1^2,i} \quad \text{Analysis of variance} \quad [13:07]$$

$$F_{k-1,N-2-k} = \frac{V(M_{0K})/(k-1)}{V_{0.1,1^2,i}/(N-2-k)} = \frac{[V(M_{0K})](N-2-k)}{[V_{0.1,1^2}-V(M_{0K})](k-1)} \quad [13:08]$$

Testing periods which are fractional introduce no complication if proper allowance is made for the decrease in N because certain X_0 values

are not used.

The columns of Table XIII A in order provide:

Column 1	Year
Column 2	Number of births
Column 3	$X_{0\Delta 1,1}{}^2$ as given by [13:03]
Column 4	$X_{0.1,1}{}^2$, residual deviations to be tested for periodicity
Column 5	Column 4 values opposite *a* are in the same phase and thus constitute one class. Values opposite *b* constitute the second class when the period being investigated is two years.
Columns 6, 7,...,21	are similar for periods of 3, 4,...,18 years.

At the feet of columns 5, 6,...,21 are recorded, in order,

$$\eta^2,\ F_{(k-1)(N-2-k)},\ P \text{ corresponding, and } \epsilon^2.$$

We first investigate trend. X_0 = number of births per year. X_1 = date, and for convenience we let $x_1 = X_1 - 1930$. We find:

$M_0 = 2050140$, and $V_0 = 4617948 \times 10^4$

$$X_{0\Delta 1} = 2050140 + 30929\ x_1$$

$$r_{01}^2 = .75956$$

To test the significance of the deviation of 30929 from zero we compute

$$F_{1,19} = \frac{V_0\ r_{01}^2\ (N-2)}{V_0\ (1-r_{01}^2)} = 60.022$$

which yields a very small *P*. We next fit a

quadric regression line, obtaining

$$X_{0\Delta 1,1^2} = 2118995 + 30929\ x_1 - 1877.9\ x_1^2 \qquad [13:09]$$

$$r^2_{0\Delta 1,1^2} = .84113$$

$$F_{1,18} = \frac{[V_0\ r^2_{0\Delta 1,1^2} - V_0\ r^2_{01}]\ (N-3)}{V_0\ (1-r^2_{0\Delta 1,1^2})} = 9.2419$$

yielding $P = .048$. Having odds of 952 to 48 that -1877.9 is not a chance deviation from zero we take [13:09] as the trend line and compute residuals

$$X_{0.1,1^2} = X_0 - X_{0\Delta 1,1^2}$$

as recorded in column 4 of Table XIII A.

We give herewith a cubic regression line, not as a part of the procedure of determining periodicity, but because it is interesting to compare such a line with the evidence available as to periodicity.

$$X_{0\Delta 1,1^2,1^3} = 2118994 + 11149\ x_1$$
$$- 1877.8\ x_1^2 + 300.61\ x_1^3$$

$$r^2_{0\Delta 1,1^2,1^3} = .89917$$

$$F_{1,17} = \frac{[V_0\ r^2_{0\Delta 1,1^2,1^3} - V_0\ r^2_{0\Delta 1,1^2}]\ (N-4)}{V_0\ (1-r_{0\Delta 1,1^2,1^3})} = 9.7856$$

yielding $P = .045$. This establishes the non-chance nature of the coefficient 300.61. We will later note that we do not establish periodicity

with the certainty that we here establish the fitness of a cubic regression line.

To test for the existence of a two-year period the residuals of column four are to be grouped into two classes, *a* and *b* as indicated in column 5. The *a* class has 11 measures and yields a mean, $M_{0\,a}$, and the 10 measures of class *b* yield a mean, $M_{0\,b}$. Weighting these means 11 and 10 respectively and computing their variance, we obtain $V(M_{0\,2})$, the subscript 2 being *k*, the number of classes involved. Analyzing variance as indicated in [13:07] and employing [13:08], we have: $F_{1.17} = .073$, yielding $P = .825$, which, being $> .5$, yields no evidence whatever that a period of two years exists. Similar computations for $T=3$, $T=4$, ..., $T=18$ have been made and are as recorded at the feet of the appropriate columns of the table. When $T = 16$, P is .152. Thus the odds are about 5 to 1 that a period in the neighborhood of 16 years exists. The F for this case is an $F_{15.3}$. We have but 3 degrees of freedom available in the error variance. Should a real period of 16 years exist it is not surprising that data covering but 21 years is unable to establish it with satisfactory certainty.

The advantages of employing $(1-P)$ in lieu of η^2, in plotting and interpreting a periodogram wherein the number of degrees of freedom is small is exemplified in Chart XIII II.

To a degree the advantages of $(1-P)$ are also inherent in ϵ^2, as is indicated by the periodogram curve based upon ϵ^2. The derivation of Chapter XI, Section 5, of ϵ^2 gave the relationship

$$\epsilon^2 = 1 - \frac{N-1}{N-k}(1-\eta^2) \qquad \text{See [11:117]}$$

but this was postulated upon using a zero order regression equation [13:01]. When a linear trend

is taken out the relationship is

$$\epsilon^2 = 1 - \frac{N-2}{N-1-k}(1 - \eta^2) \quad . \; . \; . \; . [13:10]$$

If a quadric trend is removed, we have

$$\epsilon^2 = 1 - \frac{N-3}{N-2-k}(1 - \eta^2) \quad . \; . \; . \; [13:11]$$

and so forth for the removal of trends of higher parabolic order. The values of ϵ^2 given in the table were computed by [13:11]. The median ϵ^2 value of the table is .012 which differs but slightly from .000, the value to be expected if there is no period in the data.

The periodogram analysis here discussed may be thought of as terminating the problem only in case no period is discovered. If the analysis does establish one or more periods, their further specification could be accomplished by fitting, say, a Fourier series, by a method which is appropriate when the time span of the data is not an exact multiple of the period, or periods.

SECTION 2. LEAD AND LAG IN TIME SERIES

The attempt is frequently made to predict the fluctuations of one time series from those of another. Let us consider two situations: Except for the aberrations which operate like chance *(a)* time series A and time series B fluctuate concomitantly, and *(b)* time series A tends to lag behind time series B by a constant time interval.

In the *(a)* situation the correlation between the two series may be of interest in a historical study and it may even be of interest in a prediction study if it is easy to secure the information for one of the series and difficult for the

other. Thus, if relatively easy to get steel prices neither lead nor lag behind hard to get heavy industry prices, it obviously may be of advantage to use the first to predict the second.

Knowledge of the correlation maintaining in a (*b*) situation is generally even more useful for the fluctuations of the leading variable may be used to predict the future fluctuations of the lagging variable, for example, rainfall in June might be most useful in predicting harvest in August. The three chief statistical issues of this problem are (1) by what amount of time does the predictand variable lag behind the predictor, (2) what is the correlation between them, and (3) what are the standard errors, or the fiducial limits, in the time lag measure and in the correlation measure. The third issue has, as yet, no adequate solution. A sequential analysis based upon correlated data or the employment of subsamples seems called for, but these are beyond the scope of this text. An analysis yielding a solution to the first two issues will be illustrated in connection with the data of Table XIII B.

Let X_0 be the predictand, X_1 the predictor, and X_2 the time variable. There is probably some trend in the X_0 series and the same or a different trend in the X_1 series. We can take out the trends involved by the method of the preceding Section and correlate the residual variables, pairing measures corresponding to the same moment of time and again pairing them with some time lag. This process might wholly or partially eliminate the measure of lag that concerns us. Suppose the trends are $X_{0\Delta 2} = f(X_2)$ and $X_{1\Delta 2} = f(X_2+A)$ in in which the function of X_2 in the first equa- is identical with the function of X_2+A in the second. Clearly A is exactly the measure of lag which we wish, but it has been removed in the

taking out of the trends and it will not again reveal itself by a study of the correlations, with various lags, between the residual measures $X_{0.2}$ and $X_{1.2}$. There is promise in the study of lag in an approach which requires that $X_{0\Delta 2}$ and $X_{1\Delta 2}$ be identical functions, — the first of X_2 and the second of X_2+A, if we first express the predictand and the predictor variables in comparable units. Comparability of measuring units is generally involved in any study of lag.

An obvious way to attempt to secure comparability is to deal with standard scores, as used in [13:13]. This approach may be expected to be most serviceable in connection with variables for which we have no indubitable zero points.

In the case of prices we have such a point and equal relative changes in prices are in general equally important and meaningful. In this case the equivalence of ratios method, [13:15], is fruitful. A still better procedure in this case of an indubitable zero point is to let X_0 and X_2 be the logarithms of the gross measures. Though this logarithmic transformation is appropriate to the "all foods" and the "beverages and chocolate" price indexes of Table XIII B, we will follow the more common practice of taking these price ratios as they stand for the X_0 and X_1 variables.

We pose the question, — are "all foods" retail prices a bellweather for "beverages and chocolate" retail prices and, if so, by what period of time do the latter lag and what is the correlation between them?

Pairing X_0 and X_2 items as in Table XIII B is a pairing with zero lag and it provides a sample of 84 paired measures. Pairing February X_0 with January X_1, March X_0 with February X_1, etc. is the one month lag pairing and the number in the sample is 83. Pairing March X_0 with January X_1, etc. provides the 82 cases in the two month lag

TABLE XIII B
RETAIL PRICES, 1929-35

U.S. Department of Labor, Bureau of Labor Statistics, Serial no. R 384, Date, 1938.

51 large U.S. Cities: Average 1923-15 = 100.
X_0 = "beverages and chocolate", and X_1 = "all foods."

	1929		1931		1933		1935	
	X_0	X_1	X_0	X_1	X_0	X_1	X_0	X_1
Jan.	110.7	102.7	90.2	89.2	71.1	62.6	73.5	77.4
Feb.	110.8	102.3	89.4	86.0	69.5	60.1	73.3	79.7
Mar.	110.9	101.4	87.3	85.1	68.5	59.8	72.3	79.7
Apr.	111.0	100.8	84.0	83.9	68.4	60.1	71.4	81.6
May	110.8	102.4	82.4	82.6	67.7	62.5	70.8	81.4
Jun.	110.5	103.7	82.0	80.6	67.3	64.9	70.4	81.7
Jul.	110.6	106.5	81.1	80.7	67.4	71.0	69.9	79.9
Aug.	110.4	108.1	81.1	80.9	67.7	72.0	69.3	79.6
Sep.	110.2	108.0	81.0	80.6	67.8	72.0	68.4	80.0
Oct.	110.1	107.6	80.4	79.9	68.4	71.2	68.0	80.2
Nov.	108.9	106.7	79.9	78.2	68.4	70.8	67.8	81.0
Dec.	105.3	105.7	79.4	76.2	68.0	69.7	67.6	82.2

	X_0	X_1	X_0	X_1	X_0	X_1
	1930		1932		1934	
Jan.	101.3	104.6	78.4	72.8	68.7	70.6
Feb.	99.1	103.4	77.4	70.5	69.6	72.6
Mar.	97.9	102.0	76.9	70.7	70.6	72.6
Apr.	97.2	103.3	76.0	70.3	71.4	72.2
May	96.1	102.6	74.9	68.5	72.1	73.0
Jun.	95.6	101.2	74.4	67.6	72.2	73.2
Jul.	95.7	97.5	74.2	68.3	72.1	73.5
Aug.	95.0	96.6	73.7	67.1	72.4	75.2
Sep.	93.8	98.3	74.6	66.7	72.8	76.8
Oct.	92.9	97.8	74.5	66.3	73.1	75.8
Nov.	92.2	95.2	73.8	65.6	73.0	75.2
Dec.	91.9	92.1	72.8	64.7	73.3	74.6

series, etc. The correlations resulting from these different pairings yield a series whose maximum can be found by fitting a parabola, see [7:23], and the location of this maximum, i.e., the desired period of lag, is given by [7:24].

This process may be laborious so the following short-cuts are noted: The difference correlation formula, [10:52] shows that for V_0 and V_1 constant (and they are nearly constant in going from the sample of 84 to that of 83 to that of 82, etc.) the correlation is directly dependent upon the variance of the (X_0-X_1) differences. Further, the easy to compute (see [6:68]) average deviation of these differences may be substituted for the variance of the differences to give an order of magnitude.

Thus, first compute differences (X_0-X_1) for a succession of lags, then compute the average deviation for each of these sets, stopping when a minimum is obviously passed. For each lag in the neighborhood of this minimum compute the three values, V_0, V_1, $V(X_0-X_1)$, employed in the correlation formula [10:52], and compute the requisite correlations.

Performing these steps upon the data of Table XIII B yields the information given in Table XIII C.

TABLE XIII C

CONSTANTS FOUND FOR VARIOUS LAGS IN "BEVERAGES AND CHOCOLATE"

	0 lag	1 mo. lag	2 mo. lag	3 mo. lag	4 mo. lag
N	84	83	82	81	80
$A.D.$ of (X_0-X_1)	5.006	4.759	4.752	4.805	4.986
V_1	197.250	199.634	202.031	204.491	206.950
V_0	211.259	203.619	195.404	186.962	177.950
$V(X_0-X_1)$	36.897	35.101	33.571	34.588	36.017
r	.9103	.9131	.9156	.9125	.9092

The quadric best fitting the first four points

(0 lag, .9103; 1 mo. lag, .9131; 2 mo. lag, .9156; 3 mo. lag, .9125) is given by [7:23]. It has a mode, as given by [7:21] and [7:24], at 1.808 months, revealing a lag in the retail prices of "beverages and chocolate" upon the retail prices of "all foods" of 54 days. The correlation with this lag, as given by [7:25], is .915. This completes the computation.

Serviceable standard errors could be gotten by repeating the previous process a number of times, using the data for other years, and obtaining a distribution of lag measures and also a distribution of correlation coefficients.

SECTION 3. IMPOSED CONDITIONS

Limits. A phenomenon common to the physical, biological and social sciences is that some magnitude approaches a limit as time, or age, increases. The typical growth curves of biology have an adult limiting value which they approach asymptotically as age increases. A chemical reaction may approach a state of equilibrium in a very similar manner. A bond price may gradually approach parity, or some less readily defined value determined by general economic conditions, as time passes. Not uncommonly the forgetting and the learning curves that the psychologist encounters are of the same type. In all these cases the anticipated statistical problem is first, to determine the nature of the complete growth curve from a study of complete or nearly complete antecedent phenomena, second, to fit a curve of the established type to data for which the time sequence is incomplete and third, by extrapolation secure an estimate of the limiting value.

For example, let us say that a study of the height growth curves of a substantial number of human males who have been followed from infancy

to adulthood reveals that the phenomena of increase in height follow a certain pattern, or type, and second that we have a record of growth in height of a particular boy. Then, third, we fit the type curve to the individual record and find the maximum, or adult value of the fitted curve. This we take as the estimate of the prospective adult height of the individual. There are a number of more or less difficult statistical problems involved in each of the three steps; (1) the determination of the type curve; (2) the fitting of it to the individual record; and (3) not the determination of the adult value which is very simple but the determination of its standard error (or better of its entire distribution form) which is the essential measure of credibility to be attached to the limiting, or adult, value just determined.

The algebraic solution of these problems is beyond the scope of this elementary work, but the use of graphic devices will serve as a good introduction to the sundry issues and hazards which are present.

Subordination. In medical practice certain syndromes have been found to be characteristic of certain diseases. Thus if a doctor suspects disease α he looks for symptoms A, B, C, and D, whereas if he suspects disease β he looks for symptoms B, E, F, and G. If the symptoms are qualitative, that is definitely either present or absent, very adequate contingency methods employing punched or notched cards with machine or manual sorting may be employed. These have proven very serviceable in limiting and refining diagnoses. When the symptoms are quantitative, as for example would be hours of sleep, metabolic rate, evidence of lassitude, and especially when the symptoms are in part qualitative and in part quantitative, the determination of an adequate procedure is much more involved. The problem is

common to many fields. For example the adaptability of crop to terrain is a complex function both of quantitative phenomena like rainfall and soil ingredients and of qualitative phenomena like presence or absence of a freezing temperature, a cultivation practice, etc. Though the full handling of this question is beyond the scope of this text the approach to it is through the statistics of contingency, multiple correlation and factor analysis.

SECTION 4. COMPARABLE MEASURES *

For two differently derived quantitative measures to be comparable it is logically necessary that, except for a chance factor, they measure the same underlying phenomena. It frequently happens in connection with social phenomena that a first score is in large part a measure of the same thing as a second differently derived score, as, e.g., are two intelligence test scores. Both tests are designated by their authors "intelligence" tests and both probably measure in large part the same underlying function, but in lesser part each measures a capacity found in the one but not in the other. Equating the scores of two such tests violates the logical requirements of equatable scores, but nevertheless it has been done frequently and with an outcome that has seemed serviceable. The matter of how similar two functions should be before they are equated is a problem that we leave to economists, sociologists and psychologists concerned with particular issues. We here assume that except for chance factors in each of X_1 and X_2 the underlying function measures are identical.

The equivalence of successive percentile method: If scores x_1 show a monotonic increase with increase in the underlying function and similarly for scores X_2 and if the chance factors in each

* A more detailed discussion with illustrations is given in Truman L. Kelley, STATISTICAL METHOD (1923), Chapter VI.

are negligible, then clearly the same percentile scores on X_1 and X_2 are equivalent. A table can be drawn up giving in neighboring columns the successive percentiles, $P_{.01}$, $P_{.02}, \ldots P_{.99}$ for X_1 and for X_2 and these pairs are equivalent scores. Or, in general

$$ {}_1P_p \doteqdot {}_2P_p \quad \text{Equivalent percentiles} \qquad [13:12] $$

This method does not require that the form of distribution of the X_1 scores be the same as that of the X_2 scores. In general the method is not serviceable when X_1 and X_2 have quite different reliabilities; as, e.g., would be the case if X_1 is an intelligence score on a 60 minute test and X_2 a score on a 10 minute test.

The standard score method: This method derives from the practices of Francis Galton (who, however, used medians and quartile deviations) and it asserts that standard scores as defined in [8:23] of equal magnitude are equivalent. In addition to the requirement that the underlying functions are the same, this procedure requires that $M_1 \doteqdot M_2$; that $\sigma_1 \doteqdot \sigma_2$; that the form of distribution of X_1 is the same as that of X_2; and that the relative chance errors in X_1 and X_2 are equal, namely that $\sigma_{e_1}/\sigma_1 = \sigma_{e_2}/\sigma_2$. Thus $X_1 \doteqdot X_2$ when

$$ \frac{X_1-M_1}{\sigma_1} = \frac{X_2-M_2}{\sigma_2} \quad \text{Equivalent standard scores} \qquad [13:13] $$

The estimated true standard score method: When the relative chance errors in X_1 and X_2 are unequal an improved standard score method based upon estimates of population statistics is to be preferred. In this instance it is assumed that the underlying functions are, except for chance, the same; and, if these underlying true measures are designated X_ω and X_τ; that $M_\omega \doteqdot M_\tau$; that

$\sigma_\omega \Leftrightarrow \sigma_\tau$; and that the forms of distribution of X_ω and of X_τ are the same. Then $X_\omega \Leftrightarrow X_\tau$ when

$$\frac{X_\omega - M_\omega}{\sigma_\omega} = \frac{X_\tau - M_\tau}{\sigma_\tau}$$

The sample estimates of the true statistics are $M_\omega = M_1$; $\sigma_\omega = \sigma_1\sqrt{r_1}$; $(X_\omega - M_\omega) = (X_1 - M_1)\ r_1$; and similarly for the second variable, so that $X_1 \Leftrightarrow X_2$ when

$$\frac{(X_1-M_1)\sqrt{r_1}}{\sigma_1} = \frac{(X_2-M_2)\sqrt{r_2}}{\sigma_2} \qquad \text{Equivalent estimated true standard scores} \qquad [13:14]$$

The ratio method: A common procedure in economics is to consider certain price indexes or ratios to be equivalent. Economic considerations generally justify the assumption of a common zero point, or that zero price for one commodity is equivalent to zero price for a second. If there is justification for equating another point, such as $M_1 \Leftrightarrow M_2$, or $B_1 \Leftrightarrow B_2$, in which B_1 and B_2 are some justified par, maturity values, or the like, and if the distribution of X_1 is of the same form as that of X_2, then $X_1 \Leftrightarrow X_2$ when

$$\frac{X_1}{B_1} = \frac{X_2}{B_2} \qquad \text{Equivalent ratios} \qquad [13:15]$$

SECTION 5. QUOTIENTS

There are numberless situations where some quotient X/Y is given as the crucial terminal statistic. One of the sounder trends in the study of variability and of distribution is the employment of the variance ratio. In education

and psychology, intelligence quotients, accomplishment quotients, etc. have been widely used. These procedures have very different merit because the quotients have very different properties. The variance ratio is a very special quotient in that the presumptive form of distribution of the numerator is known, as is that of the denominator; the two variances are uncorrelated; and the form of distribution of the quotient is known and available in tables. The general quotient, X/Y, is far less tractable in its algebraic handling than the variance ratio. The variance error of X/Y is generally unknown and is always difficult to compute, the difficulty in some cases being insurmountable. Of course the determination of the complete distribution of X/Y is still more difficult.

The accomplishment-quotient, which has been defined as an achievement-age divided by a mental age, has been used by some psychologists and educators for years without attaching a standard error, or other measure of variability, to it. With such limitation the quotient is not a creditable statistic. Investigators should endeavor so to set their problems that it is not needed. Economists have been equally free in the use of quotients having unknown standard errors.

The urge to employ a quotient resides in a number of facts such as that the quotient concept as applied to measures having no standard errors, such as the abstract numbers of mathematics, is well understood by everyone with elementary school training; the quotient frequently eliminates from the picture an irrelevant unit of measure; the quotient may reveal a stability not otherwise apparent, as e.g., does the intelligence quotient in a way and to a degree not immediately obvious in the mental age or gross intelligence test score.

The real understanding of a quotient based

upon observational data involves the first step of understanding the meaning of the numerator and of the denominator separately and the meaning of what is meant by division, and a second step of appreciating the trustworthiness to be attached to numerator, to denominator, and finally to the quotient. For practically all quotients the first step is simple, but for only a few is this true of the second step.

Consider two quotients *(a)* an accomplishment quotient which equals an achievement age divided by a mental age and *(b)* a variance ratio $V_{a.m}/V_{a\Delta m}$ in which $V_{a.m}$ is the variance of that part of achievement age which is independent of mental age and $V_{a\Delta m}$ is the variance of that part of achievement age which is dependent upon mental age. These two quotients bear upon the same fundamental issue and answer the same fundamental question. For the first quotient the first step in understanding is simple, but the second is so involved that it is usually entirely neglected. For the second, the first step in understanding is somewhat more difficult, but the second quite simple in that one can avail himself of the known form of distribution of the variance ratio.

Should fundamental findings be reached by variance ratios, but reported if necessary on account of the level of understanding of the audience by such quotients as accomplishment quotients the net outcome might be substantially accurate and informative. Failing this procedure, the reader of an article reporting quotients must have the idea that the writer is as little aware of the error as is the reader.

The standard error of a quotient is infrequently reported in the experimental literature, though a fairly simple formula giving it is available. To a first approximation (i.e., an approximation involving terms of the order $1/N$, but neglecting those of order $1/N^2$ and higher)

the variance error of the ratio X/Y is

$$V\left(\frac{X}{Y}\right) = \frac{X^2}{Y^2}\left(\frac{V_x}{X^2} - 2\frac{\sigma_x\sigma_y r_{xy}}{X\,Y} + \frac{V_y}{Y^2}\right) \quad \text{Variance error of a quotient} \quad [13:16]$$

There are certain shortcomings in the use of this formula. X and Y must be positive. The formula only applies when the quantity σ_y/Y is sufficiently small that neglect of powers of this quantity beyond the first will constitute no material inaccuracy. The distribution entire of X/Y is generally non-normal so that the interpretation of the critical ratio of the quotient divided by its standard error by means of the probabilities given in a normal distribution will not be precise. Finally the correlation between X and Y is required and there may be difficulties in obtaining this. With all its limitations a far wider use of this formula to give the standard error of a quotient would be a great improvement over the common practice of entirely neglecting the matter.

The quotient

$$Q = \frac{r_{12}^2}{r_{13}^2} = \frac{V_{1\Delta 2}}{V_{1\Delta 3}}$$

may be called a quotient of determination coefficients. It is indicative of the relative value of x_2 and x_3 as predictors of x_1. Accordingly, a significant excess of this ratio over 1.00 indicates that x_2 is the better predictor. The variance of Q as derived by Arthur C. Hoffman (in an unpublished paper) is

$$V_Q = \frac{4Q^2}{N}\left[\frac{1}{r_{12}^2} + \frac{1}{r_{13}^2} - \frac{2r_{23}}{r_{12}r_{13}} + \frac{2r_{13}r_{23}}{r_{12}} + \frac{2r_{12}r_{23}}{r_{13}} - r_{23}^{\,2} - 3\right] \quad [13:16a]$$

If N is, say, greater than 25, it would seem safe to interpret the critical ratio $\frac{Q}{\sigma_Q}$ in terms of the probabilities of the normal distribution.

SECTION 6. THE MOST RELIABLE WEIGHTED AVERAGE OF INDEPENDENT MEASURES

If the measures to be averaged, X_1, X_2, X_3, ..., have known variance errors, we may write each of them as equal to a true measure plus an error, thus:

$$X_1 = X_\omega + e_1;\ X_2 = X_\gamma + e_2;\ X_3 = X_\delta + e_3;\ \text{etc.}$$

If these measures are averaged, weighting them w_1, w_2, w_3, etc., we have:

$$M = \frac{w_1X_\omega\ w_2X_\gamma + w_2X_\delta + \dots}{w_1 + w_2 + w_3 + \dots} + \frac{w_1e_1 + w_2e_2 + w_3e_3 + \dots}{w_1 + w_2 + w_3 + \dots} \qquad [13:17]$$

The variance error of M is equal to the variance of the second term of the right-hand member of [13:17]. The various e's are uncorrelated as they are error functions. Knowing the variances of these several e's, which in harmony with earlier notation we designate $V_{1.\omega}$, $V_{2.\gamma}$, $V_{3.\delta}$, etc., we desire so to determine the w's that the variance error of M is minimal. We will first solve this problem for the case of two variables and will let

$$p = \frac{w_1}{w_1 + w_2}; \qquad q = 1 - p = \frac{w_2}{w_1 + w_2}$$

$$M = pX_1 + qX_2$$

$$V_m = p^2V_{1.\omega} + q^2V_{2.\gamma} = [p^2(V_{1.\omega} + V_{2.\gamma}) - 2pV_{2.\gamma}$$

$$+ \frac{V_{2.\gamma}^2}{V_{1.\omega} + V_{2.\gamma}}] + \frac{V_{1.\omega}\ V_{2.\gamma}}{V_{1.\omega} + V_{2.\gamma}}$$

The [] term is a perfect square and is the only

term containing p, so the entire function takes its smallest value when the [] = 0. Solving,

$$\frac{p}{q} = \frac{w_1}{w_2} = \frac{\dfrac{1}{V_{1.\omega}}}{\dfrac{1}{V_{2.\gamma}}} \quad . \ . \ . \ . \qquad [13:18]$$

The optimal weights are thus the inverse of the variance errors.

If we have k-variables, k-1 of which are combined (using either optimal weights, or non-optimal weights) and treated as a single variable, we can combine this with the k'th variable in the optimal manner, and the relative weight of the k'th variable, by [13:18], will be as the inverse of its variance error. We accordingly have for the optimal average of any number of variables:

$$M = \frac{\dfrac{1}{V_{1.\omega}}X_1 + \dfrac{1}{V_{2.\gamma}}X_2 + \dfrac{1}{V_{3.\delta}}X_3 + \ldots}{\dfrac{1}{V_{1.\omega}} + \dfrac{1}{V_{2.\gamma}} + \dfrac{1}{V_{3.\delta}} + \ldots} \qquad [13:19]$$

Most reliable weighted average

The most reliable weighted average of measures whose errors are independent is obtained by making the weights equal to, or proportional to, the inverse of their variance errors.

SECTION 7. FITTING OF CURVES TO OBSERVATIONS

Two broad categories are readily distinguishable. They are the fitting curves to frequency distributions and the fitting of regression lines In the first case goodness-of-fit is attested by the degree of agreement between observed frequen-

cies in the classes employed and the theoretical frequencies, or frequencies in the case of the fitted curve, in the classes. In the second case goodness-of-fit is attested by the smallness of the quantitative deviations from the regression line of the observed values.

We shall here consider the fitting of frequency distributions by the Pearson method. An alternative procedure developed by Charlier (See Rietz, op. cit., ch. VII) will not be expounded, but it has special advantages when it is desirable to express a distribution as a cumulative function of normal distributions.

The Pearson system (Pearson 1894, 1895, 1901, 1902 Cont., 1914; Elderton 1927; Craig 1936; Rietz 1924, Ch. VII by H. C. Carver) of unimodal, including anti-modal, frequency distributions is very comprehensive in that it covers the broad range of types shown on Charts XIII III and XIII IV. It employs two, three, or four parameters only, in the definition of its curves, and these are determined by the method of moments. This method establishes the equality of the moments of the fitted curve to those of the data up to the number needed, e.g., four,—mean, variance, μ_3, and μ_4,—in the case of a curve with four constants.

Further, since these moments are linear functions of the frequencies in the classes, the exact number of degrees of freedom represented by the variance of the differences of the observed frequencies in the classes and the theoretical frequencies, or those given by the corresponding areas under the fitted curve, is $(k\text{-par})$,—the number of classes employed minus the number of parameters. This enables a precise χ^2 for a test of goodness-of-fit.

The type of curve is a function of the moments. The first moment, in all cases, merely locates the curve along the x-axis. The second

moment, in all cases, merely measures variability, or, we can say, merely reflects the units in which x has been measured. Thus the important differences in type depend upon the third and fourth moments. Accordingly, all the different Pearson types can be represented by regions in a plane with dimensions μ_3/σ^3 (= $\sqrt{\text{Pearson's } \beta_1}$ = Carver's α_2) and μ_4/σ^4 (= Pearson's β_2 = Carver's α_4). Rhind (in a chart used by Pearson) has charted all regions, using as axes β_1 and β_2 [7:12] and [7:03]. We herewith give a chart of this sort, Chart XIII III.

Charts VII I, VII II, and VII III, in Chapter VII, use Pearson notation and illustrate the range of curves covered by different Pearson types.

Two criteria used by Pearson are κ_1 and κ_2 defined as follows in terms of β_1, β_2, and Craig's δ:

$$\kappa_1 = 2\beta_2 - 3\beta_1 - 6 = \frac{3\delta(\beta_1+4)}{2-\delta} \quad . \; . \; . \; . \; . \quad [13:20]$$

$$\kappa_2 = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)} = \frac{\beta_1}{4\delta(2+\delta)} \qquad [13:21]$$

Craig gives a chart of similar purpose having dimensions β_1 and δ.

$$\delta = \frac{2\beta_2-3\beta_1-6}{\beta_2+3} \qquad \text{Craig's } \delta \qquad [13:22]$$

This δ is the same as Pearson's α (See 1914, Part I, p. lxi). The use of Craig's δ, i.e., Pearson's α, leads to simplifications in equations and in charts. We also reproduce, with permission Craig's Chart,-Chart XIII IV herewith.

CHART XIII III

Diagram of Types of Frequency Distribution

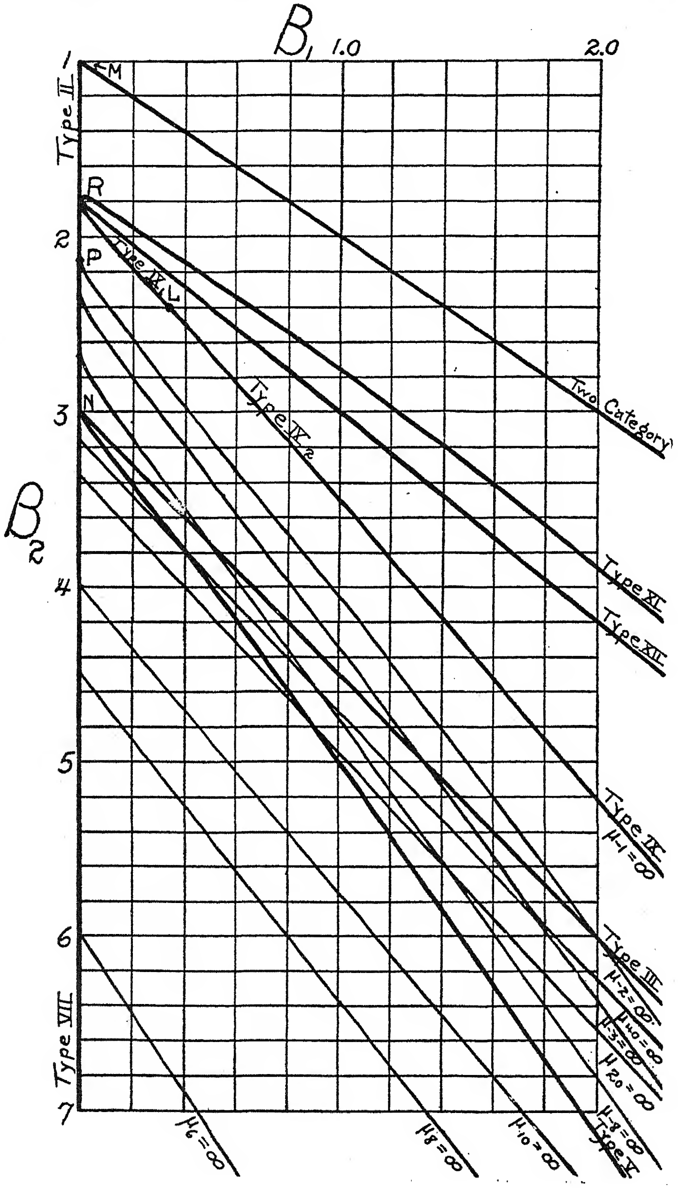
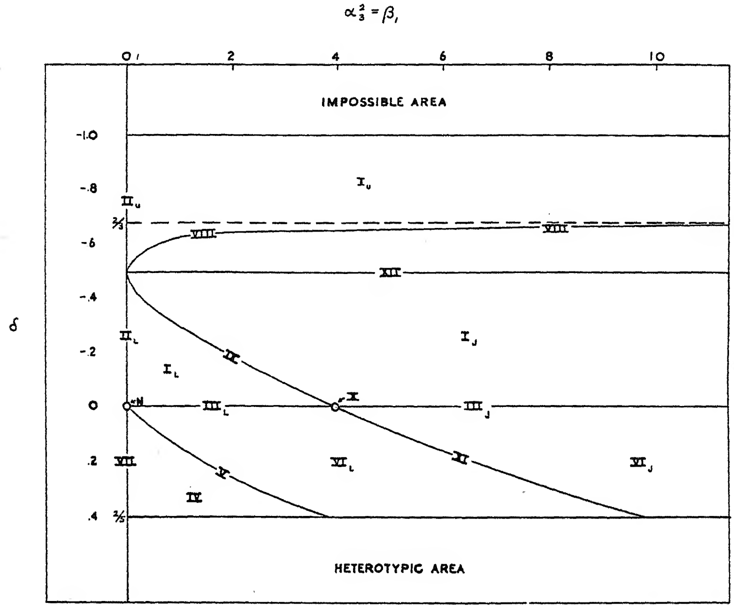


CHART XIII IV

THE (α^2, δ) CHART FOR THE PEARSON SYSTEM OF FREQUENCY CURVES

(The subscript L refers to bell-shaped curves)



On Chart XIII III are lines indicating where certain moments become infinite. This, of course, refers to the moments of the fitted curve only for any positive moment obtained from finite data cannot be infinite. The lines indicating infinite negative moments

$$\mu_{-n} = \frac{\Sigma X^{-n}}{N} \quad (X \text{ a deviation from a boundary of the curve}) \qquad [13:23]$$

also refer to the fitted curve, though an infinite negative moment is a possibility with finite data. The bearing of infinite positive and neg-

ative moments upon the stability of distributions has been discussed by the writer in *Statistical Method*, (1923). It is especially noteworthy that Type III is the only general type for which all positive moments are finite, and that Type V is the only one for which all negative moments are finite, and that the limiting distribution, represented by the intersection of the Type III and Type V lines, namely the normal distribution, is the only one having the stability that attaches to finite positive and negative moments of all orders.

The location on Craig's diagram of certain types and also certain properties of these types are given in Tables XIII D and XIII E, herewith.

TABLE XIII D

PEARSON TYPES INVOLVING THREE PARAMETERS

TYPE		
Two category	$\delta=-1$	Lower limit in δ of possible distributions. *U*-shaped curves
XII	$\delta=-.5$	*J*-shaped
III	$\delta=0$	Upper limit in δ for which all positive moments $<\infty$. Λ-shaped for $\beta_1<4$. *J*-shaped for $\beta_1>4$.
	$\delta=.4$	Upper limit in δ for which $\mu_8<\infty$.
	$\delta=1$	Upper limit in δ for which $\mu_5<\infty$.
	$\delta=2$	Upper limit in δ for which $\mu_4<\infty$.
II and VII	$\beta_1=0$	All symmetrical distributions
II	$-1<\delta<-.5$	*U*-shaped; $-.5<\delta<0$ Λ-shaped
VII	$0<\delta<2$	Λ-shaped.
VIII, IX, and XI	$\beta_1(2+3\delta)=(2+4\delta)^2(2+\delta)$ . . . . . [13:24]	
VIII	$-1<\delta<-.5$; transition between *U*- and *J*-shaped	

TABLE XIII D
(CONTINUED)

Type		
IX	$-.5<\delta<0$; transition between J- and $\wedge$-shaped	
XI	$0<\delta<2$; transition between J- and $\wedge$-shaped	
V	$\beta_1 = 4\delta(2+\delta)$	[13:25]

TABLE XIII E

TYPES INVOLVING TWO PARAMETERS

TYPE			
M	$\beta_1=0$,	$\delta=-1$.	Mendelian distribution (equal frequencies in two discrete classes).
R	$\beta_1=0$,	$\delta=-.5$.	Rectangular distribution.
N	$\beta_1=0$,	$\delta = 0$.	Normal distribution.
X or E	$\beta_1=4$,	$\delta = 0$.	Exponential distribution.
L	$\beta_1=.32$,	$\delta=-.4$.	Straight line distribution.
P	$\beta_1=0$,	$\delta=-1/3$.	Parabolic distribution.

Let the general equation covering all the Pearson curve types be

$$y = f(x) \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad [13:26]$$

for which the differential equation is

$$\frac{dy}{y\ dx} = \frac{a - x}{b_0+b_1x+b_2x^2} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \quad [13:27]$$

Written in terms of α_3 and δ (Craig's) and valid except when $\delta = -.5$ this equation is

$$\frac{dy}{y\,dx} = \frac{\dfrac{-\alpha_3}{2(1+2\delta)} - x}{\dfrac{2+\delta}{2(1+2\delta)} + \dfrac{\alpha_3}{2(1+2\delta)}x + \dfrac{\delta}{2(1+2\delta)}x^2} \qquad [13:27a]$$

This covers an extensive class of unimodal, or Λ-shaped, of *J*-shaped, and of antimodal, or *U*-shaped curves. Some of these are unlimited in both directions, some limited in one direction, and some limited in both directions, depending upon the specific values of the parameters a, b_0, b_1, b_2. The unimodal requirement limits the numerator of the right hand member of the differential equation to x to the first power only. The limitation of the denominator to x to the second power is arbitrary, but sufficiently general to encompass a very wide variety of distributions.

The curve may be fitted by the method of moments. This assures that the moments of the curve up to the requisite number agree with the moments of the data. In general it requires the first four moments, M, V, μ_3, μ_4 to determine the four parameters a, b_0, b_1, b_2, Having these, the integration of [13:27] yields the equation of distribution [13:26].

It simplifies matters to deal with standard scores, so following Carver (See Rietz, 1924, Ch. VII) and Craig (1936) we let

$$\alpha_n = \frac{\Sigma(X-M)^n}{N\sigma^n} \quad n\text{-th moment in terms of standard scores} \qquad [13:28]$$

The following recursion formula for moments holds:

$$\alpha_n a + n\alpha_{n-1}b_0 + (n+1)\alpha_n b_1 + (n+2)\alpha_{n+1}b_2 + \alpha_{n+1} \qquad [13:29]$$

Setting n successively equal to 0, 1, 2, 3 leads

to four condition equations from which we derive the following, valid except when $\delta = -.5$:

$$2(1+2\delta)a = -\alpha_3 \quad . \; . \; . \; . \; . \; . \; . \; . \quad [13:30]$$

$$2(1+2\delta)b_0 = 2+\delta \quad . \; . \; . \; . \; . \; . \; . \; . \quad [13:31]$$

$$2(1+2\delta)b_1 = \alpha_3 \quad . \; . \; . \; . \; . \; . \; . \; . \quad [13:32]$$

$$2(1+2\delta)b_2 = \delta \quad . \; . \; . \; . \; . \; . \; . \; . \quad [13:33]$$

For the general case we note that δ depends upon μ_4, μ_3, V, and M, and that α_3 depends upon μ_3, V, and M, so that the four parameters a, b_0, b_1, b_2 are determined from the first four moments. One further parameter y_0 is dependent upon N, the number of cases in the sample, or is arbitrary. The general types are called four-parameter distributions. In all cases a simple way to determine y_0 is to choose an arbitrary value, y_0', compute the frequencies in all classes with a grouping as fine as the precision of the data justifies, sum these frequencies and multiply this sum by such a number λ that the product equals N, or the arbitrarily chosen value, and then determine y_0 from [13:34].

$$y_0 = \lambda\, y_0' \qquad \text{The ordinate at the origin} \qquad [13:34]$$

To determine the Pearson [illegible] type curve that fits given data one can loc[illegible] [illegible]ate on Chart XIII IV the [illegible] point given by his data and note the [illegible] type region, or, if near a line, the type line in which or near which the point lies. What constitutes "near a type line" is considered in probability terms in *Pearson's Tables*, Part I. Great simplification occurs whenever three rather than four parameters can be used, or when two rather than three or four suffice.

Fitting the most important two-parameter curves. In addition to the normal distribution, the equation and fitting of which have already been given, these are M, R, P, L, and E.

The Mendelian distribution: The equation of this takes the non-useful form $y=y_0(1-\frac{x^2}{V})^{-1}$. The origin is at the mean, the limits are $-\sigma$, and σ, and $y_0 \to 0$. However one does not need the equation to investigate any problems of goodness-of-fit and the like which may arise.

The rectangular distribution: When $N=1$; $M=0$; range from $-\sigma\sqrt{3}$ to $\sigma\sqrt{3}$, the equation is

$$y = \frac{1}{2\sigma\sqrt{3}} \quad \text{Unit rectangular distribution} \qquad [13:35]$$

The parabolic distribution: When $N=1$; $M=0$; range from $-\sigma\sqrt{5}$ to $\sigma\sqrt{5}$, the equation is

$$y = \frac{3}{20\,\sigma^3\sqrt{5}}(5V-x^2) \quad \text{Unit parabolic distribution} \qquad [13:36]$$

The straight line distribution: When $N=1$; $M=2\sqrt{2}\,\sigma$; origin at the left end at which point $y=0$; range from 0 to $3\sqrt{2}\,\sigma$, the equation is

$$y = \frac{X}{9V} \quad \text{Unit straight line distribution} \qquad [13:37]$$

The exponential distribution: When $N=1$; $M=\sigma$; origin at the left end, at which point $y=e$; range from 0 to ∞, the equation is

$$y = e^{\frac{-x}{\sigma}} \quad \text{Unit exponential distribution} \qquad [13:38]$$

The fitting of the three-parameter distributions to data is not difficult.

The two-category distribution: $\delta=-1$. As with the Mendelian distribution the equation of the curve involves the indeterminate feature $0 \times \infty$. All distributions in which the frequen-

cies fall into two discrete categories are here represented. Letting the scores attaching to these categories be 0 and 1, and letting the proportions in them by q and p, then all the moments are functions of p, as given in Table IX A, and the distribution is completely defined. An a priori value of p is necessary if any test of goodness-of-fit is involved.

Type XII: $\delta=-.5$, When $N=1$; $M=0$; $\alpha_3=\mu_3/\sigma^3$; range from $-\sigma(\sqrt{3+\beta_1}-\alpha_3)$ to $\sigma(\sqrt{3+\beta_1}+\alpha_3)$ the equation is

$$y = y_0 \left[\frac{\sigma(\sqrt{3+\beta_1}+\alpha_3) - x}{\sigma(\sqrt{3+\beta_1}-\alpha_3) + x}\right]^{\frac{\alpha_3}{\sqrt{3+\beta_1}}} \qquad \text{Type XII} \; [13:39]$$

If y_0 is determined so that the total area = 1,

$$y_0 = \frac{1}{2\sigma\sqrt{3+\beta_1}\,\Gamma\left[\frac{\alpha_3}{\sqrt{3+\beta_1}}+1\right]\Gamma\left[1-\frac{\alpha_3}{\sqrt{3+\beta_1}}\right]} \qquad [13:40]$$

Ordinate at the point $x=\sigma\alpha_3$

To five decimal places the gamma function values may be gotten from Table XIV A.

As a sample we give herewith the Type XII equation corresponding to the ($\delta=-.5$, $\beta_1=1$) point, when μ_3 is positive, $M=0$, $\sigma=1$, $N=1$, and the range is from -1 to 3.

$$y = \frac{1}{2\pi}\left(\frac{3-x}{1+x}\right)^{\frac{1}{2}}$$

Unit Type XII distribution when $\beta_1=1$ and μ_3 is positive

This is a twisted *J*-shaped curve, i.e., one in which the low end of the *J* turns down a little.

Type III: $\delta=0$. When $N=1$; origin at the mode; range from $-a$ to ∞, the equation is

$$y = y_0 \ (1+\frac{X}{a})^{A-1} \exp \frac{-2X}{\alpha_3\sigma} \qquad \text{Type III} \qquad [13:41]$$

in which X is a deviation from the mode, and

$$Mo = M - .5\ \alpha_3\sigma \quad \ldots\ldots\ldots\ldots \quad [13:42]$$

$$\alpha_3 = \frac{\mu_3}{\sigma^3} \quad \ldots\ldots\ldots\ldots \quad [13:43]$$

$$a = \frac{2\sigma}{\alpha_3} - \frac{\alpha_3}{2\sigma} \quad \ldots\ldots\ldots\ldots \quad [13:44]$$

$$A = \frac{4}{\beta_1} \quad \ldots\ldots\ldots\ldots \quad [13:45]$$

$$y_0 = \frac{(A-1)}{a}\left[\frac{(A-1)^{A-1}}{e^{A-1}\Gamma A}\right] \qquad \text{Ordinate at the mode} \qquad [13:46]$$

in which the reciprocal of the [] term is given by [14:50]. When $\beta_1 = 0$ this reduces to the normal distribution and when $\beta_1 = 4$ it becomes the exponential. Herewith is given as an example the Type III equation when μ_3 is positive, $\beta_1 = 1$, $\sigma = 1$, $N = 1$, $Mo = 0$, the range then being from -1.5 to ∞.

$$y = \frac{1}{3}\, e^{-2X-3}(3+2X)^3 \qquad \text{Unit Type III having } \beta_1 = 0 \qquad [13:47]$$

The distribution of variances from samples drawn from a normal population is of Type III. This type has been found to be widely descriptive of phenomena. Areas, ordinates and derivatives for Type III distributions have been tabled by Salvosa (1930).

Type II. $\beta_1 = 0$; $-1 < \delta < 0$ (i.e., $1 < \beta_2 < 3$). When $N = 1$; origin at the mean; range from $-a$ to a, the equation is

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^m \quad \text{Type II distribution} \qquad [13:48]$$

$$M = Mo, \text{ or anti-mode} = Mdn = 0 \qquad [13:49]$$

$$m = \frac{5\beta_2 - 9}{2(3-\beta_2)} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \qquad [13:50]$$

$$a^2 = \frac{2\,V\,\beta_2}{3 - \beta_2} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \qquad [13:51]$$

$$y_0 = \frac{\Gamma(m - 1.5)}{a\sqrt{\pi}\,\Gamma(m + 1)} \quad \text{Ordinate at the mean} \qquad [13:52]$$

When $\delta = -1$ (i.e., $\beta_2 = 1$), a Mendelian distribution of equal frequencies in two classes results,—see Chart VII I, Type M.

When $\delta = -.5$ (i.e., $\beta_2 < 1.8$), m is negative and a U-shaped curve results,—see Chart VII I, Type II-U.

When $-.5 < \delta < 0$ (i.e., $1.8 < \beta_2 < 3$), m is positive and a platykurtic Λ-shaped curve results,— Chart VII I, Type II-i.

When $\delta = -1/3$ (i.e., $\beta_2 = 2\frac{1}{7}$), the distribution is parabolic,—see Chart VII I, Type P.

When $\delta = 0$ (i.e., $\beta_2 = 3$), a normal distribution results.

Type VII. $\beta_1 = 0$; $0 < \delta < 2$ (i.e., $3 < \beta_2 < \infty$). When $N = 1$; origin at the mean; range from $-\infty$ to ∞, the equation is

$$y = y_0 \left(1 + \frac{x^2}{a^2}\right)^{-m} \quad \text{Type VII distribution} \qquad [13:53]$$

$$M = Mo = Mdn = 0 \quad \ldots\ldots\ldots\ldots \quad [13:54]$$

$$m = \frac{5\beta_2 - 9}{2(\beta_2 - 3)} \quad \ldots\ldots\ldots\ldots \quad [13:55]$$

$$a^2 = \frac{2\ V\ \beta_2}{\beta_2 - 3} \quad \ldots\ldots\ldots\ldots \quad [13:56]$$

$$y_0 = \frac{\Gamma m}{a\sqrt{\pi}\ \Gamma(m-1.5)} \quad \text{Ordinate at the mean} \quad [13:57]$$

For illustration, see Chart VII I, curve A.

Types VIII, IX, and XI. $\beta_1 = (2+3\delta) = (2+4\delta)^2 (2+\delta)$ as given in [13:24]. In the equations for these three types occurs an exponent, m, which is found by solving the accompanying cubic,

$$m^3(\beta_1-4) + m^2(9\beta_1-12) + 24m\beta_1 + 16\beta_1 = 0 \quad [13:58]$$

The solution may follow the steps indicated in Chapter XIV, Section 7. A plot of [13:58] shows that, in addition to an extraneous branch, there is a branch passing through the following points: $\beta_1 = \infty$, $m = -1$; $\beta_1 = 0$, $m = 0$; $\beta_1 = 4$, $m = \pm\infty$; $\beta_1 = \infty$, $m = -4$. The region between the first two points yields Type VIII distributions and herein $-1 < m < 0$. Between the second and third point yields Type IX distributions and herein $0 < m < \infty$. Between the third and fourth point yields Type XI distributions and herein $-\infty < m < -4$. The distributions represented by special points upon this cubic are R, L, and E, as noted in Table XIII E. A convenient form of the equation of the curve for Types VIII and IX is

$$y = y_0 \left(1 + \frac{X}{a}\right)^m \quad \text{Types VIII and IX} \quad [13:59]$$

and for Type XI,

$$y = y_0 \ X^m \quad . \quad . \quad . \quad . \quad \text{Type XI} \quad . \quad . \quad . \quad . \qquad [13:60]$$

The separate and specific features of these three types will not be considered.

Type VIII. $-1 < m < 0$; $0 < \beta_1 < \infty$; and also $-1 < \delta < -.5$. A limited range transition type between *U*- and *J*-shaped curves. The equation as written has $N = 1$, and the origin at the right boundary when $\mu_3 > 0$ and at the left boundary when $\mu_3 < 0$. When $\mu_3 > 0$ the range is from $-a$ to 0 and when $\mu_3 < 0$ it is from 0 to $-a$ ($-a$ being a positive magnitude).

$$a = \pm\, \sigma(2+m)\sqrt{\frac{3+m}{1+m}} \quad \text{The sign of } a \text{ is the same as that of } \mu_3 \qquad [13:61]$$

$$y_0 = \frac{1+m}{\pm\, a} \quad \text{Sign so chosen that } y_0 > 0. \text{ Ordinate at boundary.} \qquad [13:62]$$

$$M = \frac{-a}{2+m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad [13:63]$$

Type IX. $0 < m < \infty$; $0 < \beta_1 < 4$; and also $-5 < \delta < 0$. A limited range transition type between *J*- and Λ-shaped curves. As written the origin is at the left boundary when $\mu_3 > 0$ and at the right boundary when $\mu_3 < 0$. The range is from 0 to $-a$ when $\mu_3 > 0$ and from $-a$ to 0 when $\mu_3 < 0$.

$$a = \pm\, \sigma(2+m)\sqrt{\frac{3+m}{1+m}} \quad \text{The sign of } a \text{ is opposite that of } \mu_3 \qquad [13:61a]$$

y_0 is given by [13:62]

M is given by [13:63]

Type XI. $-\infty < m < -4$; $4 < \beta_1 < \infty$; and also $0 < \delta < 2$. The curve is unlimited at one end and is a transition type between *U*- and Λ-shaped curves. When $\mu_3 > 0$ (and the scale of measurement can always be so chosen that this is true) the equation as written has $N = 1$, and the origin at 0, which is the distance a to the left of the left boundary of the curve. The range is from a to ∞.

$$a = -\sigma\ (2+m)\sqrt{\frac{3+m}{1+m}} \quad (\text{When } \mu_3 > 0) \ . \ . \ . \qquad [13:64]$$

$$y_0 = -(1+m)a^{-(1+m)} \quad \text{Ordinate at } X=0, \text{ a point outside the range} \qquad [13:65]$$

$$M = \frac{a(1+m)}{2+m} \quad \text{Mean measured from the origin} \qquad [13:66]$$

The mean is $-a/(2+m)$ to the right of the left boundary a.

Type V. $\beta_1 = 4\delta(2+\delta)$. A leptokurtic curve of limited range in one direction. When $N = 1$; origin at the left boundary when $\mu_3 > 0$; range from 0 to ∞; the equation is

$$y = y_0\ X^{-p}\ e^{\frac{-\gamma}{X}} \ . \ . \ . \ . \ . \ . \ . \ . \ \text{Type V} \qquad [13:67]$$

$$p = 4 + \frac{8+4\sqrt{\beta_1+4}}{\beta_1} \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \qquad [13:68]$$

$$\gamma = (p-2)\sigma\ \sqrt{p-3} \quad \text{Sign of } \gamma \text{ to be the same as that of } \mu_3 \qquad [13:69]$$

$$y_0 = \frac{\gamma^{p-1}}{\Gamma(p-1)} \quad \text{This constant is not an ordinate at any real point} \qquad [13:70]$$

$$M = \frac{\gamma}{p-2} \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \qquad [13:71]$$

$$Mo = M - \frac{2\gamma}{p(p-2)} \quad \ldots\ldots\ldots\ldots \quad [13:72]$$

As a sample we write down the Type V equation when $N=1$; $\sigma=1$; $\alpha_3=1.5$; origin = 0. Then $p=16$; $\gamma=14\sqrt{13}$; $y_0=(14\sqrt{13})^{15}/\,\Gamma\,15$; $M=\sqrt{13}$; $M_0=\frac{7}{8}\sqrt{13}$, and

$$y = \frac{(14\sqrt{13})^{15}}{\Gamma 15} X^{-16} \exp \frac{-14\sqrt{13}}{X}$$

The four parameter Types I, IV, and VI. For the fitting of these we refer to original memoirs, or to Elderton, op. cit.

Type I is given by points in the Craig Chart for which $\delta<0$. The general equation, with origin at the mode, is

$$y = y_0 \left(1+\frac{X}{a_1}\right)^{m_1} \left(1-\frac{X}{a_2}\right)^{m_2} \qquad \text{Type I} \qquad [13:73]$$

Type VI is given by points in the region below the line $\delta=0$ and above the line $\beta_1=4\delta(2+\delta)$. The general equation, with origin at a before the start of the curve, is

$$y = y_0 (X-a)^{m_1} \quad X^{-m_2} \qquad \text{Type VI} \qquad [13:74]$$

Type IV is given by points in the region below the line $\beta_1=4\delta(2+\delta)$ and to the right of the line $\beta_1=0$. This is the only region in which the range is unlimited in both directions. The general equation, with origin at $\nu\, a/r$ beyond the mean is

$$y = y_0 \left(1+\frac{X^2}{a^2}\right)^{-m} \exp \left[-\nu \tan^{-1}\left(\frac{X}{a}\right)\right] \qquad [13:75]$$

Type IV

SECTION 8. THE VARIANCE ERROR OF A STATISTIC DERIVED FROM ONE HAVING A KNOWN VARIANCE ERROR

Let K be the statistic having the known variance error V_k and let D be the statistic derived from K, whose variance error, V_d, is desired. In general σ_k/K is a magnitude having a factor $1/\sqrt{N}$, $1/\sqrt{N-1}$, or the like, so that, if N is at all appreciable the square and higher powers of σ_k/K are negligible in comparison with σ_k/K. In this case, but only in this case, the methods here given suffice.

The binomial expansion method: Students with no knowledge of the calculus can use this method and it has a surprisingly wide field of applicability. Using the bar to indicate mean values when many samples are taken, and letting d be the error in the sample value of D and k the error in the sample value of K, we have $D=\bar{D}+d$, and $K=\bar{K}+k$, so that

$$\bar{D}+d = f(\bar{K} + k)$$

Expressing $f(\bar{K}+k)$ as a binomial with positive, negative, integral, or fractional exponents, expanding and keeping the terms of zero and first degree only in k, as may be done if higher degree terms than the first in (k/K) are negligible in comparison with the first, we obtain

$$\bar{D} + d = \phi(\bar{K}) + k\,\Psi(\bar{K})$$

the variables when many samples are taken are d and k, so we have

$$V_d = [\Psi(\bar{K})]^2 V_k \qquad [13:76]$$

The variance of a derived statistic as a function of the variance of the given statistic (binomial expansion method)

To illustrate this method let us derive the variance of the standard deviation knowing the variance of the variance.

$$\bar{\sigma} + s = (\bar{V}+v)^{\frac{1}{2}} = \bar{V}^{\frac{1}{2}} + \frac{v}{2\bar{V}^{\frac{1}{2}}}$$

Computing the variance of the right and left hand members we have

$$V_{\sigma} = \frac{V_v}{4\bar{V}} \doteq \frac{V_v}{4V} \qquad \text{Variance error of } \sigma \text{ in case } N \text{ is ample} \qquad [13:77]$$

Substitution of statistical deviations for calculus differentials method. The degree of accuracy in this procedure is the same as in that just given. Using the same notation as before it is assumed that the calculus ratio $\frac{dD}{dK}$ is essentially equal to the statistical ratio d/k. Given $D=f(K)$ we obtain the derivative

$$dD = f'(K)\ dK$$

and substitute for the differentials, writing

$$d = f'(K)\ k,$$

thence immediately

$$V_d = [f'(K)]^2\, V_k \qquad \text{The variance of a derived statistic (statistical deviations for differentials method)} \qquad [13:78]$$

To illustrate in the case of V_{σ}

$$\sigma = V^{\frac{1}{2}}$$

the differential equation is

$$d\sigma = \frac{dV}{2V^{\frac{1}{2}}}$$

substituting statistical deviations, noting that

$\sigma = \bar{\sigma} + s$, and that $V = \bar{V} + v$

$$s = \frac{v}{2V^{\frac{1}{2}}}$$

so that, computing the variance of both members,

$$V_\sigma = V_s = \frac{V_v}{4V} \qquad \text{Variance error of } \sigma \text{ in case } N \text{ is ample} \qquad [13:79]$$

Expanding by a Taylor (or Maclaurin) series method. The degree of accuracy depends upon the number of terms of the series kept. When kept to the f' term the accuracy is the same as in the preceding methods. Using the same notation as before,

$$D = \overline{D} + d = f(\overline{K} + k)$$

$$= f(\overline{K}) + k\ f'\ (\overline{K}) + \frac{k^2}{2} f''(\overline{K}) + \ . \ . \ .$$

Keeping the expansion to two terms and taking the variance of both members,

$$V_d = [f'(\overline{K})]^2\ V_k \qquad \text{The variance of a derived statistic (Taylor's series method)} \qquad [13:80]$$

To illustrate in the case of V_σ:

$$\sigma = \overline{\sigma} + s = (\overline{V} + v)^{\frac{1}{2}}$$

$$= V^{\frac{1}{2}} + v\,\frac{1}{2\overline{V}^{\frac{1}{2}}}$$

and taking the variance of both members

$$V_\sigma = \frac{V_v}{4V} \doteq \frac{V_v}{4V} \qquad \text{The variance error of } \sigma \text{ in case } N \text{ is ample} \qquad [13:81]$$

Logarithmic differentials method. This is a special case of the substitution of statistical

deviations for differentials method, which is very convenient when the derived statistic is a function of products or quotients of other statistics. Let

$$D = K^a L^b M^c \quad \ldots\ldots\ldots \quad [13:82]$$

wherein the variances and covariances of the variables K, L, and M are known. Taking the logarithm of [13:82] we have

$$\log D = a \log K + b \log L + c \log M \quad [13:83]$$

Taking logarithmic differentials

$$\frac{dD}{D} = a\frac{dK}{K} + b\frac{dL}{L} + c\frac{dM}{M}$$

Substituting statistical deviations for differentials, squaring, summing, and dividing by the number of samples to get variances, yields

$$\frac{V_d}{D^2} = a^2\frac{V_k}{K^2} + b^2\frac{V_1}{L^2} + c^2\frac{V_m}{M^2} + 2ab\frac{c_{kl}}{K\,L} + 2ac\frac{c_{km}}{K\,M} + 2bc\frac{c_{lm}}{L\,M} \quad [13:84]$$

Variance of a derived statistic (logarithmic differentials method)

An illustration of this method is given in Section 9.

SECTION 9. THE VARIANCE ERROR OF A COEFFICIENT OF CORRELATION CORRECTED FOR ATTENUATION

We shall employ logarithmic differentials in the computation of this variance error. The coefficient of correlation corrected for attenuation is most commonly derived from three observed values: (1) the correlation between the scores upon two tests; (2) the correlation between similar halves of the first test, i.e., the half

test reliability, and (3) the half test reliability for the second test. We designate the variables involved as follows:

x_3 and x_5 are the similar halves of x_1, the first measure.

x_4 and x_6 are the similar halves of x_2, the second measure.

$x_1 = x_3 + x_5$; $x_2 = x_4 + x_6$; $V_3 = V_5$; $V_4 = V_6$; x_3 and x_5 correlate approximately equally with other variables, as do also x_4 and x_6.

For convenience we employ such units that $V_3 = V_4 = 1$.

$V_1 = 2+2r_{35}$; $V_2 = 2+2r_{46}$; With these units covariances are: $c_{35} = r_{35}$; $c_{46} = r_{46}$;

$c_{12} = 2\sqrt{1+r_{35}}\sqrt{1+r_{46}}\ r_{12} = c_{34}+c_{36}+c_{54}+c_{56} =$

$4c_{34}$; $c_{31}=c_{51}=1+r_{35}$; $c_{42}=c_{62}=1+r_{46}$; $c_{32}=c_{52}$

$=c_{41}=c_{61}=2c_{34}=c_{12}/2$

In the formula for the coefficient corrected for attenuation

$$r_{\omega\gamma} = \frac{r_{12}}{\sqrt{r_1}\sqrt{r_2}}$$

the r_1 and r_2 are not observed values but Spearman-Brown formula [11:10] stepped up values. Before developing a formula for its variance error we must express $r_{\omega\gamma}$ in terms of the observed items, thus

$$r_{\omega\gamma} = \frac{r_{12}}{\sqrt{\dfrac{2r_{35}}{1+r_{35}}}\sqrt{\dfrac{2r_{46}}{1+r_{46}}}} \quad . \; . \; . \; . \; . \quad [13:85]$$

$$2\,r_{\omega\gamma} = r_{12}\,(r_{35})^{-\frac{1}{2}}(1+r_{35})^{\frac{1}{2}}(r_{46})^{-\frac{1}{2}}(1+r_{46})^{\frac{1}{2}} \quad [13:86]$$

Taking logarithmic differentials of [13:86], substituting statistical deviations for differentials, and computing the variance, we obtain

$$\frac{4V_{r_{\omega\gamma}}}{r^2_{\omega\gamma}} = \frac{V_{r_{12}}}{r^2_{12}} + \frac{V_{r_{35}}}{4r^2_{35}} + \frac{V_{r_{46}}}{4r^2_{46}} + \frac{V_{r_{35}}}{4(1+r_{35})^2}$$

$$+ \frac{V_{r_{46}}}{4(1+r_{46})^2} - \frac{C_{r_{12}r_{35}}}{r_{12}r_{35}} + \frac{C_{r_{12}r_{35}}}{r_{12}(1+r_{35})} - \frac{C_{r_{12}r_{46}}}{r_{12}r_{46}}$$

$$+ \frac{C_{r_{12}r_{46}}}{r_{12}(1+r_{46})} - \frac{V_{r_{35}}}{2r_{35}(1+r_{35})} + \frac{C_{r_{35}r_{46}}}{2r_{35}r_{46}} - \frac{C_{r_{35}r_{46}}}{2r_{35}(1+r_{46})} \qquad [13:87]$$

$$- \frac{C_{r_{35}r_{46}}}{2(1+r_{35})_{46}} + \frac{C_{r_{35}r_{46}}}{2(1+r_{35})(1+r_{46})} - \frac{V_{r_{46}}}{2r_{46}(1+r_{46})}$$

If we now substitute variances of correlation coefficients as given by [10:46] and covariances as given by [13:134] we secure, after some simplification

$$V_{r_{\omega\gamma}} = \frac{r^2_{\omega\gamma}}{4(N-2)}\left(4r^2_{\omega\gamma} + \frac{4}{r^2_{12}} + \frac{1}{r^2_{35}} + \frac{1}{r^2_{46}} - \frac{4}{r_{35}} - \frac{4}{r_{46}} - 2\right) \qquad [13:88]^*$$

Variance error of $r_{\omega\gamma}$ when obtained via [13:85] or [13:91]

By a similar process we find that if a correction for attenuation in one variable only is

* B. Babington Smith has kindly checked this derivation.

made so that

$$r_{1\gamma} = \frac{r_{12}}{\sqrt{\dfrac{2r_{46}}{1+r_{46}}}} \qquad \text{Coefficient of correlation corrected for attenuation in one variable only} \qquad [13:89]$$

The variance error is

$$V_{r_{1\gamma}} = \frac{r_{1\gamma}^2}{N-2}\left(r_{1\gamma}^2 + \frac{1}{r_{12}^2} - \frac{1}{r_{46}} + \frac{1}{4r_{46}^2} - \frac{5}{4}\right)^* \qquad [13:90]$$

Variance error of $r_{1\gamma}$ when obtained via [13:89]

A way equally excellent to [13:85] for utilizing all the data to obtain $r_{\omega\gamma}$ is by the use of Yule's formula.

$$r_{\omega\gamma} = \frac{(r_{34}r_{36}r_{54}r_{56})^{\frac{1}{4}}}{\sqrt{r_{35}r_{46}}} \qquad \text{G. U. Yule's formula for } r_{\omega\gamma} \qquad [13:91]$$

The writer has proven that to the degree of precision given by terms involving $1/(N-2)$ the variance error of $r_{\omega\gamma}$ thus determined is the same as of $r_{\omega\gamma}$ determined by [13:85].

SECTION 10. THE EQUI-PROBABLE AND THE MEAN RANGES FOR SAMPLES OF DIFFERENT SIZE DRAWN FROM A NORMAL POPULATION

We let the scores, as deviations from the mean in terms of the population standard deviation, be designated x. From Tippett's table (1925)(see also Pearson, 1932) we can find the probability that the value x will be the largest observation for a sample of N (Tippett's tables cover samples of $N = 3, 5, 10, 20, 30, 50, 100, (100), 1000$). Let us choose a small interval (the writer

*Stanley Smith Stevens, H. L. Long, and E. R. Stabler have kindly checked this derivation.

chose $i = .2\sigma$) and designate the probability of x in the interval $(x-.5i)$ to $(x+.5i)$ f_x, as readily computed from Tippett's tables. Let us arbitrarily choose some range, ra. Let p_{x-ra} = the probability that a measure drawn at random from the normal population $< (x-ra)$. This probability is immediately available from a table of normal probability functions. Let p_x = the probability that a measure drawn at random $< x$. Then $(p_x-p_{x-ra})/p_x$ is the probability that in a sample whose largest measure is x a second measure $> (x-ra)$, and the probability that the $(N-1)$ measures other than the measure $x > (x-ra)$ is $[(p_x-p_{x-ra})/p_x]^{n-1}$. Thus the probability that the range in this sample of $N > ra$ is

$$\{1-[(p_x-p_{x-ra})/p_x]^{n-1}\}$$

The frequency of occurrence of this situation is f_x. Finally we sum $f_x\{1-[(p_x-p_{x-ra})/p_x]^{n-1}\}$ for all successive values of x to obtain the probability that the observed range will exceed ra. We can repeat this process for different values of ra until we experimentally find that value for which the probability = .5.

To illustrate in the case of N=10. With ra = 3.1σ the computation described yielded p=.46240, the probability that a range $> 3.1\,\sigma$ would be obtained from a sample of 10. Then ra = 3.00 was investigated and the probability of a range $> 3.0\,\sigma$ was found to be .51228. Linear interpolation between these values to find the range for which p = .5 gives $3.0246\,\sigma$, the equi-probable range as recorded in Table XIII F herewith. The mean range is a little greater than the equi-probable range. Its value, as given by Tippett is recorded in the last column.

TABLE XIII F

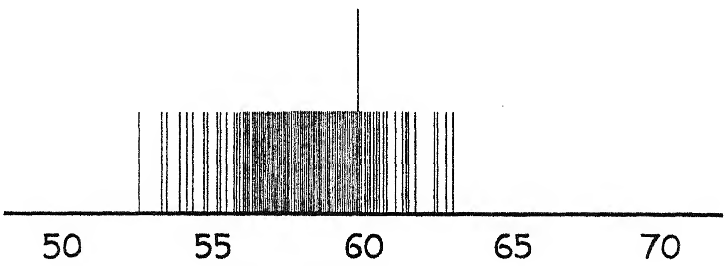
THE EQUI-PROBABLE RANGE, AND THE MEAN RANGE, OF SAMPLES DRAWN FROM A NORMAL POPULATION (in terms of the population σ)

N	EQUI-PROBABLE RANGE	MEAN RANGE (FROM TIPPETT)
6	2.46	2.5344
10	3.0246	3.0775
25	3.90	3.9306
50	4.47	4.4981
100	4.9683	5.0152
300	5.70	5.7555
1000	6.4379	6.4829
10000	7.67	
100000	8.78	

SECTION 11. THE OPTIMAL SIZE OF INTERVAL FOR HISTOGRAM OR FREQUENCY POLYGON

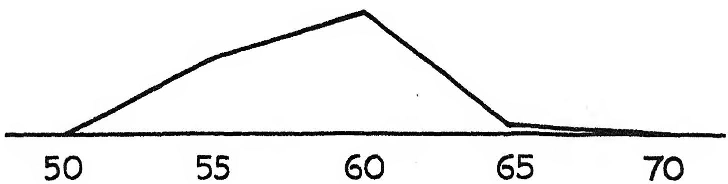
The temperature data, Table IV J, are not finely enough graduated to illustrate well the issues involved in connection with the interval for graphic portrayal, so we will deal with the heights (hypothetical) of 100 American male white adults recorded to the nearest .01 of an inch, Table XIII F. With such fine grouping, few if any classes will have a frequency greater than 1. A plot with this very fine grouping is shown in Chart XIII V.

CHART XIII V



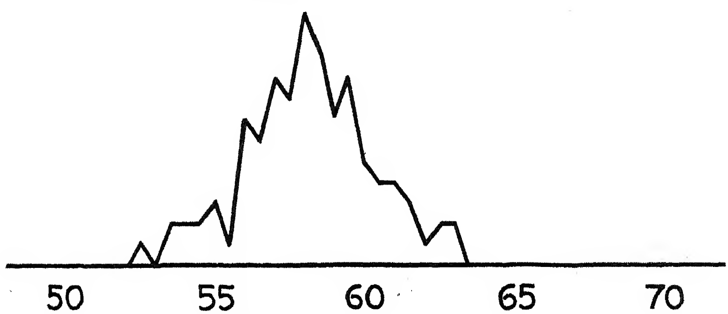
which, all will agree, is quite uninformative. If we group in intervals of 5 inches, it is as shown in Chart XIII VI, which again is not

CHART XIII VI



very informative. We clearly desire to group using some interval greater than .01 inch and less than 5 inches. If we group using .5-inch intervals the curve of Chart XIII VII results.

CHART XIII VII



To the trained statistician, aware of the fluctuations to be expected from sampling, this might be quite satisfactory. However, graphic portrayal is essentially a popular device and is supposed to convey information to the statistically initiated. If such a person sees Chart XIII VII he likely will interpret as significant all the small fluctuations from class to class. He certainly has no standard as to where to draw the line between significant and insignificant fluctuations. This is in fact a difficult line for the well-trained statistician to draw. The ori-

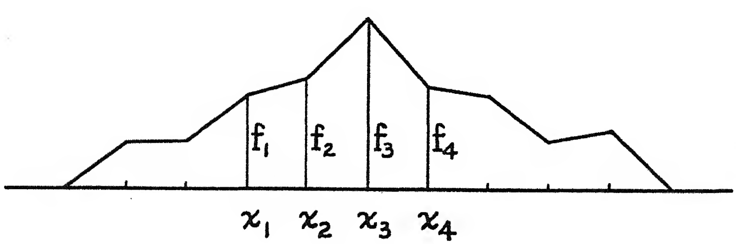
ginal measurements, we will say, are as given in Table XIII G.

TABLE XIII G

52.63	56.45	57.68	58.67	59.84
53.30	56.59	57.76	58.73	59.91
53.47	56.67	57.77	58.74	59.96
53.91	56.68	57.81	58.81	60.13
54.09	56.77	57.84	58.87	60.30
54.35	56.81	57.90	58.97	60.37
54.72	56.86	57.98	58.98	60.45
54.78	56.91	58.00	59.06	60.67
55.07	56.95	58.04	59.13	60.81
55.17	57.11	58.09	59.19	60.83
55.43	57.13	58.13	59.26	60.93
55.81	57.17	58.20	59.32	61.09
55.89	57.24	58.25	59.39	61.37
55.96	57.27	58.27	59.44	61.59
56.08	57.36	58.33	59.50	61.61
56.11	57.43	58.36	59.51	61.75
56.19	57.48	58.41	59.63	62.41
56.23	57.49	58.51	59.69	62.58
56.34	57.53	58.55	59.71	62.87
56.36	57.63	58.58	59.84	63.03

We therefore seek a rule for coarseness of grouping for graphic portrayal which will result in a frequency polygon which, probably, will not be misleading to the untutored reader who interprets the directions of the frequency fluctuations shown as meaningful. If the graph looks like that of Chart XIII VIII, we desire that the

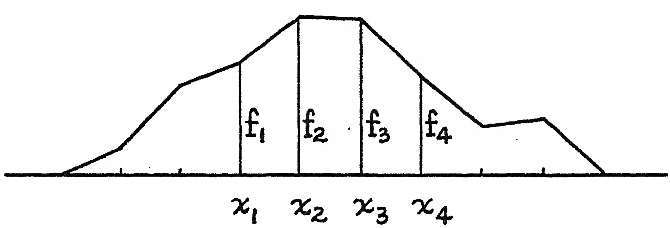
CHART XIII VIII



reader have at least an even chance of being right if he believes that in the population the frequency in class x_2 exceeds that in x_1 and falls short of that in x_3. Of all the possible comparisons between class frequencies we may say that the edict that f_3 is greater than f_2 or f_4 is the one most likely to be made, and the one above all others that we should seek to make trustworthy.

The foregoing is not quite as refined a statement of the problem as is necessary because it does not cover a case that approximates that of Chart XIII IX. Here the naive judgment is that the mode lies between x_2 and x_3. We should hope that this judgment would be one having a 50:50 or better chance of being right.

CHART XIII IX



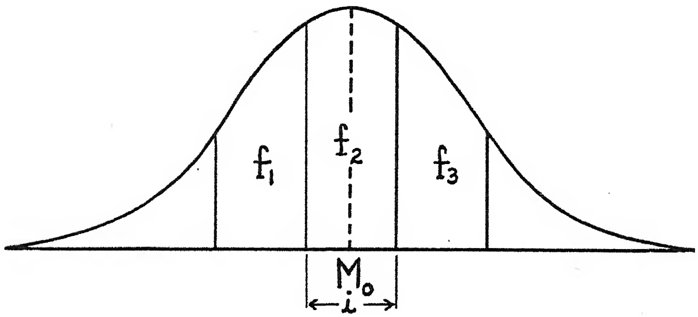
We will now state specifically the problem set: It is to determine the coarseness of grouping to employ so that the observed mode (intermediate between two class indexes if the naive judgment so places it there) of a sample of N drawn from a normal population has a probability of .5 of lying in an interval having the true mode as its class index. If we can determine this interval, then we may conclude that the mode shown in a graphic representation using this interval, of a sample drawn from a normal population, will have a probability of .5 of not being in error by more than one-half the interval used, i.e., $\text{P.E.}_{\text{Mo}} = \frac{i}{2}$, if the parent population is normal.

Except for the infrequent samplings yielding a greater frequency in a class removed two or more intervals from the true modal class than in the modal class, this statement is equivalent to the following: Given a normal population grouped in classes, with interval i for convenience expressed in terms of σ as the unit, and with the true mode as the class index of one of the classes, then the desired interval i is such that the probability is .5 that, in a sample, f_2 is greater than f_1 and also greater than f_3.

If $f_2 > f_1$ then $(f_2-f_1) > 0$, and if $f_2 > f_3$ then $(f_2-f_3) > 0$, and in a scatter diagram with axes (f_2-f_1) and (f_2-f_3) the relative frequency above the line $f_2-f_1 = 0$ is the probability that $f_2 > f_1$, and the relative frequency to the right of the line $f_2-f_3 = 0$ is the probability that $f_2 > f_3$, and the relative frequency in the upper right sector is the probability that $f_2 > f_1$ and also $> f_3$.

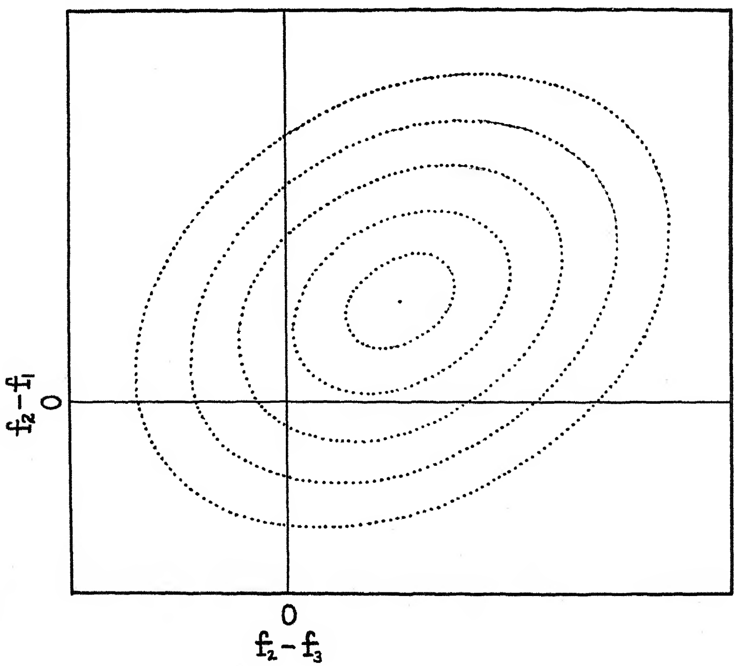
Specifically employing an interval, i, let us be given three successive classes, with sample frequencies f_1, f_2, and f_3, such that the class index of the second class coincides with the parent normal population mode. Let N = the number in the sample. Let $\tilde{f}_1$, $\tilde{f}_2$, and $\tilde{f}_3$, in which

CHART XIII X



$\tilde{f}_1 = \tilde{f}_3$, be the population frequencies in these three classes. In the population $\tilde{f}_2$ is the largest, but because of the vicissitudes of sampling, f_2 may not be the largest in the sample. The greater the interval the more likely it is that f_2 will be the largest. We shall determine the interval, associated with a given N, for which the probability is .5 that f_2 is the largest. We shall consider this the desirable interval for graphic portrayal because then, except for the small chance that some $f_4 > f_1$ or f_2 or f_3, we have at least an even chance that the sample mode will not differ by more than half an interval from the true mode. Since it would only be by rare chance that the class index of the second class would coincide with the population mode, we must consider the sample mode to be

CHART XIII XI



that obtained by smoothing the frequencies in the four classes constituting the modal group for the sample, by some such process as [7:21] and [7:24].

The correlation surface Chart XIII XI will be approximately normal if the f's are not too small. Assuming that many samplings are made, the constants of this correlation situation are readily available. They are as given herewith, in which a circumflex over a symbol indicates a population value, and M, V, and r are the mean, variance, and correlation. Also $p = \frac{f}{N}$ and $q = 1 - p$.

$$M_{f_2-f_1} = \tilde{f}_2 - \tilde{f}_1 = \tilde{f}_2 - \tilde{f}_3 = M_{f_2-f_3}$$

$$V_{f_2-f_1} = N\tilde{p}_2\tilde{q}_2 + N\tilde{p}_1\tilde{q}_1 + 2N\tilde{p}_1\tilde{p}_2 = V_{f_2-f_3}$$

$$r_{(f_2-f_1)(f_2-f_3)} = \frac{\tilde{p}_2\tilde{q}_2 + 2\tilde{p}_1\tilde{p}_2 - \tilde{p}_1^2}{\tilde{p}_2\tilde{q}_2 + \tilde{p}_1\tilde{q}_1 + 2\tilde{p}_1\tilde{p}_2}$$

In terms of the standard deviation, the point of dichotomy at which $f_2 - f_1 = 0$ is $M_{f_2-f_1}/\sigma_{f_2-f_1}$ below the mean of the distribution. Similarly for the point in the other variable at which $f_2 - f_3 \doteq 0$. It is now a simple matter to determine i by interpolating between the tables by Lee *et alia* (Pearson, 1914) giving the proportionate frequencies in upper right quadrangles for different dichotomies and different correlations. This has been done for certain values of N as reported in Table H herewith.

TABLE XIII H

THE SIZE (in σ units) OF THE INTERVAL GIVING A PROBABILITY OF .5 THAT THE FREQUENCY OF THE MEDIAN INTERVAL SHALL EXCEED THAT OF EITHER NEIGHBORING INTERVAL, IN THE CASE OF A NORMAL DISTRIBUTION; THE EXPECTED RANGE (in σ units); AND THE NUMBER OF CLASSES NECESSARY TO COVER THIS RANGE

N	i GIVING $P = .5$	Ra EXPECTED RANGE	NO. OF CLASSES $= Ra/i$
6	.90562	2.46	2.72
10	.81215	3.0246	3.724
25	.67082	3.90	5.81
50	.58140	4.47	7.69
100	.50402	4.9683	9.857
300	.40347	5.70	14.13
1000	.31226	6.4379	20.62
10000	.19490	7.67	39.4
100000	.12499	8 .78	70.

The equi-probable ranges given in the third column are those derived in the immediately preceding Section 10. The values in the fourth column, rounded off to the nearest integer, give the practical answer that we have sought, the number of classes to employ in the graphic portrayal of samples of different sizes. The information of the last column is repeated in Chapter IV, Table IV M, in a form which is simpler to use.

SECTION 12. DIRECT AND INVERSE INTERPOLATION

There are many methods of interpolation where intervals are equal. A time saving method for direct interpolation is by means of tabled Lagrangian interpolation coefficients as exemplified in Tables XV A and XV B.

There are various methods available when two-way interpolation is called for. One serviceable method is by successive one-way interpolations.

Satisfactory direct interpolation when intervals are unequal is more complicated. For one treatment of this matter, see E. T. Whittaker and G. Robinson, *The Calculus of Observations*, 1924, Ch. II.

Inverse interpolation occurs when the tabled values (unequal intervals) are used for purposes of entry and the answer sought is in terms of the variable (equally spaced) which is usually the argument. Linear inverse interpolation is simple, but if insufficiently accurate quadric inverse interpolation as later described may be used.

a is the argument for which a value t is desired or in the inverse problem t is the argument for which a is desired. $a_0<a<a_1$. The interval in the equally spaced arguments is i.

$$i = a_1 - a_0 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [13:92]$$

The fractional distance of the a_0 to a_1 interval to reach a is p.

$$p = \frac{a-a_0}{a_1-a_0} = \frac{a-a_0}{i} \quad \text{also } q = 1-p. \qquad [13:93]$$

For inverse interpolation

$$p' = \frac{t-t_0}{t_1-t_0} = \frac{t-t_0}{\Delta_0^{\mathrm{I}}} \quad \text{also } q' = 1-p'. \qquad [13:94]$$

When it is necessary to distinguish between a's and t's for different p's they are designated a_p and t_p. When it is necessary to distinguish between t's gotten by two-, three-, four-point, etc. interpolation they are designaced t^{II}, t^{III}, t^{IV}, etc. When it is necessary to distinguish between a's gotten by two-, three-, and four-point inverse interpolation they are designated

a^{-ii}, a^{-iii}, a^{-iv}.

We define symbols as in Table XIII I and as in the following defining equations.

TABLE XIII I

NOTATION FOR ARGUMENTS, TABLED ENTRIES AND DIFFERENCES.

ARGU-MENTS	TABLED ENTRIES	DIFFERENCES 1ST ORDER	2ND ORDER	3RD ORDER	4TH ORDER	5TH ORDER	6TH ORDER
a_{-2}	t_{-2}						
		Δ^{I}_{-2}					
a_{-1}	t_{-1}		Δ^{II}_{-2}				
		Δ^{I}_{-1}		Δ^{III}_{-2}			
a_0	t_0		Δ^{II}_{-1}		Δ^{IV}_{-2}		
		Δ^{I}_{0}		Δ^{III}_{-1}		Δ^{V}_{-2}	
a_1	t_1		Δ^{II}_{0}		Δ^{IV}_{-1}		Δ^{VI}_{-2}
		Δ^{I}_{1}		Δ^{III}_{0}		Δ^{V}_{-1}	
a_2	t_2		Δ^{II}_{1}		Δ^{IV}_{0}		
		Δ^{I}_{2}		Δ^{III}_{1}			
a_3	t_3		Δ^{II}_{2}				
		Δ^{I}_{3}					
a_4	t_4						

In connection witn inverse interpolation we require Δ and δ following:

$$\Delta = \frac{\Delta_0^{II}}{\Delta_0^{I}} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [13:95]$$

$$\delta = \frac{-\Delta_{-1}^{III}}{\Delta_0^{I}} + 3\Delta^2 \quad . \; . \; . \; . \; . \; . \; . \quad [13:96]$$

We shall judge of the accuracy of interpolation by a parabola of a certain degree by comparing the result obtained thereby with that obtained by using a parabola of the next higher degree. This comparison generally leads to a slight, but only slight, overstatement of the error.

The following formulas give successively closer approximations to the correct value t_p.

$$t_p^{ii} = (1-p)\,t_0 + p\,t_1 \qquad \text{Two-point, or linearly interpolated value} \qquad [13:97]$$

$$t_p^{iii} = p(1-p)(2-p)\left[\frac{1}{2p}\,t_0 + \frac{1}{1-p}\,t_1 - \frac{1}{2(2-p)}\,t_2\right] \qquad \text{Three-point, or quadric value} \qquad [13:98]$$

If $p < .5$ interpolate in the other direction using t_1, t_0, and t_{-1}.

$$t_p^{iv} = \frac{1}{6}p(1-p^2)(2-p)\left[\frac{-1}{1+p}\,t_{-1} + \frac{3}{p}\,t_0 + \frac{3}{1-p}\,t_1 - \frac{1}{2-p}\,t_2\right] \qquad \text{Four-point, or cubic value} \qquad [13:99]$$

Formulas [13:98] and [13:99] are of the forms

$$t_p^{iii} = c_0 t_0 + c_1 t_1 - c_2 t_2 \qquad [13:98a]$$

$$t_p^{iv} = -c_{-1} t_{-1} + c_0 t_0 + c_1 t_1 - c_2 t_2 \qquad [13:99a]$$

These c-coefficients are known as Lagrangian interpolation coefficients. Tables giving them for interpolation up to eleven-point and for different values of p are available. See Chapter XV, Tables XV A and XV B and references given in Chapter XV, Section 1.

A good approximation to the error in t_p^{ii} is $(t_p^{iii} - t_p^{ii})$; in t_p^{iii} it is $(t_p^{iv} - t_p^{iii})$ and in t_p^{iv} it is $(t_p^{v} - t_p^{iv})$. These errors are maximal when $p = .5$. In this case we designate them E^{ii}, E^{iii}, and E^{iv}.

$$E^{ii} = \frac{1}{8}(-t_0 + 2t_1 - t_2) \quad \text{Maximal error in linear interpolation} \qquad [13:100]$$

$$E^{iii} = \frac{1}{16}(t_{-1} - 3t_0 + 3t_1 - t_2) \qquad [13:101]$$

Maximal error in three-point interpolation

$$E^{iv} = \frac{3}{128}(t_{-2} - 4t_{-1} + 6t_0 - 4t_1 + t_2) \qquad [13:102]$$

Maximal error in four-point interpolation

For the Table XV C of the normal probability functions given in Chapter XV, Section 3, the E^{ii} and E^{iii} values recorded at the bottom of columns give these maximal linear and quadric interpolation errors as determined from values approximately halfway down the columns.

For *inverse interpolation*, i.e., the procedure for obtaining a knowing t, the following formulas hold:

$$a^{-ii} = a_0 + ip' \quad \text{Linear inverse interpolation} \quad [13:103]$$

$$a^{-iii} = a_0 + i\left(p' + \frac{\Delta p'q'}{2 + 3\Delta + \Delta^2}\right)$$

or approximately

$$a^{-iii} = a_0 + i[p' + .25p'q'\Delta(2-3\Delta)] \quad [13:104]$$

Inverse quadric interpolation

Defining the error in inverse two-point interpolation as the difference between the three-point and two-point inverse interpolation answers, and designating the maximum value of this error E^{-ii}, we have

$$E^{-ii} = .125\ i\ |\Delta| \quad \text{Maximum error in inverse linear interpolation} \quad [13:105]$$

Similarly,

$$E^{-iii} = .0625\ i\ |\delta| \quad \text{Maximum error in inverse quadric interpolation} \quad [13:106]$$

SECTION 13. THE MACHINE EXTRACTION OF SQUARE AND CUBE ROOTS

A number of excellent methods are available, so the only question is that of expedition of process.

The extraction of square root by the following procedure is very rapid: We assume a computing machine with no trick devices. It has a keyboard, a product dial or register, and a rotation counter dial or register, which latter is sometimes called the multiplier, or quotient dial. To illustrate the process let us desire the square root of 123.456.

Set 123.456 in the product dial.

Set 11. (an inspectional estimate of the answer) in the keyboard.

Divide, keeping the quotient to 4 figures (twice the number in the trial root) obtaining 11.22.

Make mental note of 11.11, the average of 11. and 11.22.

Clear the rotation counter by multiplying by 11.22, thus restoring 123.456 to the product dial.

Change the keyboard to 11.11 and divide obtaining 11.112151.

The average of this and 11.11, namely 11.11076, is the desired root correct to eight figures, which is twice as many figures as the number which is the same in the second approximation (11.11) and the quotient (11.112151).

Let us extract the cube root of 123.456. If a slide rule or brief table of cube roots, Chapter XV, Table D, can be used to get a first estimate the process will be shortened by one or two iterations. The slide rule gives 4.98.

Set 4.98 in the keyboard and square. The product dial reading is 24.8004, but this does not need to be remembered or recorded.

The carriage is returned so that a rotation will register in the unit's place.

The keyboard is cleared.

The multiplication bar is depressed once, not changing the product reading, but increasing the 4.98 by 1 to 5.98.

The keyboard is set to be identical with the product reading.

The subtraction is made changing the counter dial to 4.98 and clearing the product dial, Glance at the product dial to see that this is so.

Set the "division" lever.

Multiply by such a number, namely 4.98, as clears the rotation counter dial. The product dial now holds the cube of 4.98.

If this cube is greater than 123.456 set the "division" lever, or button, and if less set the "multiplication" lever. In the present problem this cube = 123.50599, so something must be sub-

tracted from it. We accordingly set the "division" lever.

Make sufficient addition or subtractions (in this case subtractions) to make the product dial reading = 123.456. When this is accomplished we find -.002016 (minus because the "division" lever is operating) appearing in the counter dial.

One-third of this, namely -.00672, is the correction to 4.98. Accordingly the answer is 4.979328. This is correct to twice as many figures as the trial root was correct, i.e., it is correct to the last figure. Otherwise stated 4.9793280 differs from 4.980 in the fourth significant figure so 4.9793280 is in error in the eighth significant figure.

If the value at this point is not accurate to enough decimal places repeat the process using the improved estimate of the cube root. Usually two such computations will suffice to give eight figure accuracy.

SECTION 14. OCCASIONAL FORMULAS

Since the proportion, p, of cases in a class is the frequency, f, in that class divided by N and since, under conditions of sampling in which N is made constant from sample to sample, the distributional characteristics of p are simply, those of f divided by the constant N, we can immediately derive statistics of p from those of f, as given in Chapter IX, formulas [9:02], [9:03], and [9:04]. We have $Np=f$; $N^2p^2=f^2$; etc. so that

$$V_p = \frac{pq}{N} \qquad \text{Variance of a proportion} \qquad [13:107]$$

$$U_3(p) = \frac{pq(q-p)}{N^2} \qquad [13:108]$$

Third moment of a proportion

$$U_4(p) = \frac{pq[1+3(N-2)pq]}{N^3} \qquad [13:109]$$

Occasionally one needs the regression, the correlation, or the covariance, between the frequencies, f_a, f_b, or the proportions, p_a, p_b, in two classes. Having the variances of the frequencies, formula [9:02], both regression and correlation become available when the covariance is known. We here derive the covariance, $c(f_a f_b)$, between the frequencies in two classes by first deriving the regression of one of these frequencies upon the other.

Let all classes except the two in question be considered as a single class, c. The sample frequencies are f_a, f_b, f_c, such that $f_a+f_b+f_c = N$. The true, or population frequencies are $\tilde{f}_a$, $\tilde{f}_b$, $\tilde{f}_c$, such that $\tilde{f}_a+\tilde{f}_b+\tilde{f}_c=N$. Letting Δ's represent deviation in frequency from true values we write $f_a=\tilde{f}_a+\Delta_a$ and $f_b=\tilde{f}_b+\Delta_b$. When Δ_b is some positive number of cases there must be an exactly compensating deficiency in the other two classes and in the long run the deficiency that attaches to the *a* class will bear the ratio to the deficiency that attaches to the c class that $\tilde{f}_a$ bears to $\tilde{f}_c$. That is, the regression of Δ_a upon Δ_b is given by the equation

$$\Delta_a = \frac{-\tilde{f}_a}{\tilde{f}_a+\tilde{f}_c}\Delta_b$$

This is the regression of f_a upon f_b and after making simple substitutions the regression coefficient can be written

$$b(f_a f_b) = \frac{-\tilde{p}_a}{\tilde{q}_b} \doteq \frac{-p_a}{q_b} \qquad \text{Regression of one cell frequency upon a second non-overlapping frequency} \qquad [13:110]$$

From this, utilizing the appropriate variances and standard deviations, we immediately obtain

$$r(f_a f_b) = r(p_a p_b) = -\sqrt{\frac{p_a p_b}{q_a q_b}} \qquad [13:111]$$

Correlation between non-overlapping cell frequencies, or proportions

$$c(f_a f_b) = -Np_a p_b \qquad [13:112]$$

Covariance between non-overlapping cell frequencies

$$c(p_a p_b) = \frac{-p_a p_b}{N} \qquad [13:113]$$

Covariance between non-overlapping cell proportions

The preceding results may be applied to the frequencies, or the proportions, in a two-dimensional contingency table. In case one of the frequencies is a part of the second, this second can be split into two parts, one correlating perfectly with the first frequency and the other part having the correlation [13:111]. Making allowance for this we can obtain the covariances for all the type situations that maintain in a two dimensional contingency table. From these covariances one can readily compute the covariances between proportions and also the correlations between frequencies and the correlations between proportions. We designate the marginal totals for the rows f_a, f_b, etc., those for the columns $f_{a'}$, $f_{b'}$, etc., and those for the cell lying at the intersection of the *a* row and the *a′* column $f_{aa'}$, and similarly for cells at other intersections. It is simple to establish that

$$c(f_{aa'} f_{bb'}) = -Np_{aa'} p_{bb'}$$

This being [13:112] in new notation

$$c(f_{aa'} f_b) = Np_{aa'} q_a \qquad [13:114]$$

Covariance between the frequency in a row and the frequency in a cell in that row

$$c(f_a f_{a'}) = N(p_{aa'} - p_a p_{a'})$$ Covariance between the frequency in a row and the frequency in a column [13:115]

If the first variable is quantitative, having a mean, M_1, and a deviation of row *a* from the mean, x_a, then, as established by Pearson (1913)

$$c(M_1 f_{aa'}) = p_{aa'} x_a$$ Covariance between a mean and a cell frequency in a scatter diagram [13:116]

If both variables are quantitative the true regression equation of X_1 upon X_2 is

$$\overline{X}_1 = \tilde{b}_{12} X_2 + \tilde{M}_1 - \tilde{b}_{12} \tilde{M}_2$$

This equation holds for every observed X_2 value. Thus for individual *i* in a sample of *N* we have

$$\overline{X}_{1i} = \tilde{b}_{12} X_{2i} + \tilde{M}_1 - \tilde{b}_{12} \tilde{M}_2$$

If such equations are written down for all the *N* cases in the sample, added and divided by *N*, we have, by noting that $\Sigma X_2 = NM_2$ and $\Sigma X_1 = NM_1$,

$$\overline{M}_1 = \tilde{b}_{12} M_2 + \tilde{M}_1 - \tilde{b}_{12} \tilde{M}_2$$

This establishes that the parameters in the regression of M_1 upon M_2 are the same as those in the regression of X_1 upon X_2. Thus *the regression of means is the same as the regression of variables*, and it follows that *the correlation between means is the same as that between variables when samples are randomly drawn from a single parent population.* The best approximation to these true parameters are the sample values, yielding

$$b(M_1 M_2) = b_{12} \quad \ldots\ldots \quad [13:117]$$

$$r(M_1 M_2) = r_{12} \quad \ldots\ldots \quad [13:118]$$

$$c(M_1 M_2) = \frac{1}{N} c_{12} \quad \ldots\ldots \quad [13:119]$$

If in an experimental situation, such e.g., as a sampling of public opinion, both $r(M_1M_2)$ and r_{12} can be independently calculated and when so calculated are found to be different, it is established that the successive samples have not been random samples from the same parent population.

We here compute a regression, correlation, and covariance between two variances. We consider the population distribution and take $\tilde{M}_1$ and $\tilde{M}_2$ as the arbitrary origins. We note that the mean $\tilde{x}_1^2$ for an array $= \tilde{x}_{1\Delta 2}^2 + \tilde{V}_{1.2}$. Since $\tilde{x}_{1\Delta 2} = \tilde{b}_{12}\tilde{x}_2$ this yields the following regression of x_1^2 upon $\tilde{x}_2^2$:

$$\overline{\tilde{x}_1^2} = \tilde{b}_{12}^2 \tilde{x}_2^2 + \tilde{V}_{1.2}$$

Summing for a sample of N, and dividing by N, we obtain

$$\overline{\tilde{V}}_1 = \tilde{b}_{12}^2 \, {}_1\tilde{V}_2 + \tilde{V}_{1.2} \quad \ldots\ldots\ldots \quad [13:120]$$

in which ${}_1\tilde{V}_2 = (\Sigma\tilde{x}_2^2)/N$ for the sample of N cases. It is an unbiased estimate of $\tilde{V}_2$, which fact we may express thus: ${}_1\tilde{V}_2 = \tilde{V}_2 + \overline{0}$, in which $\overline{0}$ is an unbiased estimate of 0, i.e., the mean of these for many samples approaches 0. Thus from [13:120] we get

$$\tilde{b}(V_1V_2) = \tilde{b}_{12}^2 = \tilde{r}_{12}^2 \frac{\tilde{V}_1}{\tilde{V}_2} \qquad \text{Regression of variances} \qquad [13:121]$$

$$\tilde{r}(V_1V_2) = \tilde{r}_{12}^2 \qquad \text{Correlation between variances} \qquad [13:122]$$

$$\tilde{c}(V_1V_2) = \tilde{r}_{12}^2 \sqrt{V_{V_1}} \sqrt{V_{V_2}} \qquad \text{Covariance between variances} \qquad [13:123]$$

For the best value of r_{12}^2 based upon sample data we may use [11:114]. For the best sample estimates of V_1 and V_2 we may use [6:09], and for V_{V_1} and V_{V_2} we note that $\frac{N-1}{N}V = \tilde{\bar{V}}$, and that

$$V(\frac{N-1}{N}V) = \frac{(N-1)^2}{N^2} V_V$$

so multiplying the right hand member of [6:55] by $\frac{N^2}{(N-1)^2}$ will yield the desired estimates of the true variances needed in [13:123].

Further, if we deal with situations in which N is large and in which we may assume that the parent bivariate population is normal, the preceding formulas simplify, yielding:

$$b(V_1V_2) = b_{12}^2 = r_{12}^2\frac{V_1}{V_2} \qquad [13:124]$$

N large and bivariate normal population

$$r(V_1V_2) = r_{12}^2 \qquad [13:125]$$

N large and bivariate normal population

$$c(V_1V_2) = \frac{2}{N}c_{12}^2 \qquad [13:126]$$

N large and bivariate normal population. SEE [13:148]

Further variance, correlation, and covariance formulas have been derived by Karl Pearson (1913) and are here given:

We employ the P_{ij} and the p_{ij} notation as given in Chapter XI, Section 4. That is $P_{10}=M_1$; $P_{01}=M_2$; $p_{11}=c_{12}$; $p_{20}=V_1$; $p_{02}=V_2$; etc. The variance error of any product moment, *when N is large*, that is, terms of order $1/N^2$, $1/N^3$ etc., are neglected, is given by

$$N\,V(p_{ij}) = p_{2i,2j} - p_{ij}^2 + i^2 p_{20} p_{i-1,j}^2$$

$$+ j^2 p_{02} p_{i,j-1}^2 + 2ij\,p_{11} p_{i-1,j} p_{i,j-1}$$

$$- 2i\,p_{i+1,j} p_{i-1,j} - 2j\,p_{i,j+1} p_{i,j-1} \qquad [13{:}127]$$

(N large) Variance error of any product moment from the means

The covariance between any two product moments, each involving the same two variables, is given by

$$N\,c(p_{gh} p_{ij}) = p_{g+i,\,h+j} - p_{gh} p_{ij} + g\,i\,p_{20} p_{g-1,h} p_{i-1,j}$$

$$+ h\,j\,p_{02} p_{g,h-1} p_{i,j-1} + g\,j\,p_{11} p_{g-1,h} p_{i,j-1}$$

$$+ h\,i\,p_{11} p_{g,h-1} p_{i-1,j} - i\,p_{g+1,h} p_{i-1,j}$$

$$- j\,p_{g,h+1} p_{i,j-1} - g\,p_{i+1,j} p_{g-1,h}$$

$$- h\,p_{i,j+1} p_{g,h-1} \qquad (N \text{ large}) \qquad [13{:}128]$$

Covariance of product moments from the means (two-variables)

Since formulas [13:127] and [13:128] pertain to deviations from the means, none of g, h, i, or j may = 1. Formulas applying without this restriction as given by Karl Pearson (op. cit) are [12:129] and [13:130]. As in the two preceding formulas g and i refer to the powers of a first variable and h and j to the powers of a second variable. When the origins of these variables are fixed points we have the variance formula [13:129] and the covariance formula [13:130]. In each instance N should be ample.

$$N\ V(P_{ij}) = P_{2i,2j} - P_{ij}^2 \quad . \ . \ . \ . \ . \quad [13{:}129]$$

$$N\sigma_{P_{gh}}\sigma_{P_{ij}}r_{P_{gh}P_{ij}} = Nc(P_{gh}P_{ij}) = P_{g+i,h+j} - P_{gh}P_{ij} \quad [13{:}130]$$

Special cases of the two preceding formulas in all of which terms of order less than $1/N$ are neglected, and certain further formulas are given herewith:

$$p_{22} = V_1 V_2(1+2r^2) \quad \text{Assumption of normality} \quad [13{:}131]$$

$$r(\sigma_1\sigma_2) = r_{12}^2 \quad \text{Assumption of normality} \quad [13{:}132]$$

$$p_{13} = 3\ r_{12}\ \sigma_1\ \sigma_2^3 \quad \text{Assumption of rectilinearity only)} \quad [13{:}133]$$

$$r(M_1\sigma_1) = 0 \quad \text{Assumption of symmetry} \quad [13{:}134]$$

$$r(M_1\sigma_1) = r(M_1V_1) = \sqrt{\frac{\beta_1}{\beta_2 - 1}} \quad . \ . \ . \ . \ . \ . \quad [13{:}135]$$

$$r(r_{12}\sigma_1) = \frac{r}{\sqrt{2}} \quad \text{Assumption of normality} \quad [13{:}136]$$

$$r(r_{12}\sigma_1) = \frac{r_{12}}{2(1-r_{12}^2)} \frac{(\sqrt{\beta_{2(\text{1st var.})}-1} - r_{12}^2\sqrt{\beta_{2(\text{2nd var.})}-1}}{[1-.25(\beta_{2(\text{1st})}+\beta_{2(\text{2nd})}-6\dfrac{r_{12}^2}{1-r_{12}^2}]^{\frac{1}{2}}} \quad [13{:}137]$$

$$r(r_{12}M_1) = 0 \quad \text{Assumption of normality} \quad [13{:}138]$$

Let $a = M_1 - b_{12}M_2$, the constant term in a regression equation, then

$$c(ab_{12}) = -\frac{M_2 V_1(-r_{12}^2)}{N\ V_2} \quad \text{Assumption of normal-ity)} \qquad [13:139]$$

Involving three variables we have,

$$r(\sigma_1 r_{23}) = \frac{r_{12}(r_{13}-r_{12}r_{23}) + r_{13}(r_{12}-r_{13}r_{23})}{(1-r_{23}^2)\ \sqrt{2}}$$

Assumption of normality [13:140]

$$2N\ c(r_{12}r_{13}) = [2r_{23}-r_{12}r_{13}-2\,r_{23}(r_{12}^2+r_{13}^2)$$
$$+r_{12}r_{13}(r_{12}^2+r_{13}^2+r_{23}^2)] \qquad [13:141]$$

Assumption of normality

Involving four variables we have,

$$N\ c(r_{12}r_{34}) = r_{13}r_{24}+r_{14}r_{23}-r_{12}r_{13}r_{14}$$
$$-\ r_{12}r_{23}r_{34}-r_{13}r_{23}r_{34}-r_{14}r_{24}r_{34}$$
$$+\frac{1}{2}r_{12}r_{34}(r_{13}^2+r_{14}^2+r_{23}^2+r_{24}^2)$$

Assumption of normality [13: 142]

Formulas [13:140], [13:141], and [13:142] were first given by Filon and Pearson (1898, pp. 229-311).

Romanowsky (1925) gives the following formula for the variance of a regression coefficient:

$$V(b_{12}) = \frac{V_1\ (1-r^2)}{V_2\ (N-3)} \quad \text{Assumption of normality cf. [10:41]} \qquad [13:143]$$

Karl Pearson (op.cit) gives the following vari-

ance for the constant term in a two variable regression equation:

$$V_a = (M_2^2 + V_2)\ V(b_{12}) \quad \text{Assumption of normality} \qquad [13:144]$$

The following formula holding for any form of bivariate distribution if N is large for the variance error of r was first given by Sheppard (1898).

$$N\ V(r_{12}) = r_{12}^2 \left(\frac{p_{22} - p_{11}^2}{p_{11}^2} + \frac{p_{40} - p_{02}^2}{4\ p_{20}^2} + \frac{p_{04} - p_{02}^2}{4\ p_{02}^2} \right.$$

$$+ \frac{p_{22} - p_{20}p_{02}}{2\ p_{20}p_{02}} - \frac{p_{31} - p_{11}\ p_{20}}{p_{11}p_{20}}$$

$$\left. - \frac{p_{13} - p_{11}p_{02}}{p_{11}p_{02}} \right) \ldots\ldots\ldots \quad [13:145]$$

Formulas for a number of the preceding standard errors and correlations, not involving the assumption of normality, are given by Isserlis, (1916). In another article (1918) he gives reduction formulas for higher product moments, in normal multivariate distributions such, for example, as for

$$p_{112} (= \Sigma x_1 x_2 x_3^2 / N)$$

The eight following formulas, holding for samples from a normal bivariate population, are from Wishart (1928):

$$N\bar{c}_{12} = (N-1)\tilde{c}_{12}, \text{ or in other notation } \frac{N-1}{N}\tilde{p}_{11} = c_{12} = p_{11}$$

normal bivariate population [13:146]

$$N^2V(V_1) = 2(N-1)\tilde{V}_1^2 \qquad \text{See [6:58]} \qquad [13:147]$$

$$N^2c(V_1V_2) = 2(N-1)\ \tilde{c}_{12}^2 \qquad \text{See [13:126] (normal bivariate population)} \qquad [13:148]$$

$$N^2V(c_{12}) = (N-1)(\tilde{V}_1\tilde{V}_2 + \tilde{c}_{12}^2) \qquad \text{Normal bivariate population)} \qquad [13:149]$$

$$N^2c(V_1c_{12}) = 2(N-1)\ \tilde{V}_1\ \tilde{c}_{12} \qquad \text{Normal bivariate population)} \qquad [13:150]$$

$$N^2c(V_1c_{23}) = 2(N-1)\ \tilde{c}_{12}\tilde{c}_{13} \qquad \text{Normal trivariate population)} \qquad [13:151]$$

$$N^2c(c_{12}c_{13}) = (N-1)(\tilde{V}_1\tilde{c}_{23} + \tilde{c}_{12}\tilde{c}_{13}) \qquad [13:152]$$

Normal trivariate population)

$$N^2c(c_{12}c_{34}) = (N-1)(\tilde{c}_{13}\tilde{c}_{24} + \tilde{c}_{14}\tilde{c}_{23}) \qquad [13:153]$$

Normal quadrivariate population)

Sundry uni-, bi-, and multivariate formulas for various moments and product moments, holding for small samples of any form are given by both Pepper (1929), and Feldman (1935).

SECTION 15. SEQUENTIAL ANALYSIS

This is a method wherein observations of the items in a lot are taken in succession until the cumulated evidence warrants the acceptance or rejection of the lot. This inspectional use of sequential analysis differs only in minor aspects from its use in experimentation and in quality control.

We here consider its use in connection with inspection of a product which can be classified as defective, or non-defective, and in which the number of cases in the lot is large in comparison with

the undetermined, but much smaller number, inspected. Adaptation to the inspection of a product having a graduated score as, for example, tensile strength of a piece of metal or scholastic standing of a college student is, in general, simple.

The abscissa of Chart XIII XII is the number of items inspected and the ordinate is the number of defective items. The broken line indicates the results as found after one item, two items, three items, etc. have been inspected. This sample number line is also a time line showing the amount of information given at each successive inspectional stage. All the information due to earlier inspections is inherent in the status shown by the last plotted point on the line. For any point, the ordinate divided by the abscissa gives the proportion of defective items, an unbiased estimate of the true proportion in the lot. Of course this estimate is the more reliable the more items are inspected.

The proportion of defectives found at any stage of inspection is an unbiased estimate of the proportion maintaining in the lot, but when the number of items inspected is small the estimate is relatively quite unreliable, particularly when the proportion is very high or very low.

Prior to any knowledge of the actual number of defectives found after n-inspections, the ordinate at n can be divided into three parts: an upper portion wherein the number of defectives is so great that the lot is rejected, a middle portion wherein decision is reserved, and a lower portion wherein the number is so small that the lot is accepted. This statement holds after the first few inspections, for in practical situations a final decision either way would not be made upon the basis of one or two or other small number of inspections even if all the items inspected were perfect or all imperfect. Since the division of every ordinate into these three parts is possible,

we obtain, by connecting homologous points, the three regions shown on Chart XIII XII. It is necessary to lay down specific conditions so that the lines demarking these regions may be specifically defined. Two lines, those for d_1 and d_2, which, conveniently, turn out to be straight lines have particular merit. These lines are functions of p_1, p_2, α, and β defined as follows:

p_1 is a proportion of defectives so small that a lot having this proportion is considered acceptable.

α is the risk of rejecting a lot in which $p = p_1$. This risk is generally small for it is considered highly undesirable to reject so excellent a lot.

p_2 is a proportion of defectives, greater than p_1, of such size that a lot having this proportion is considered unacceptable.

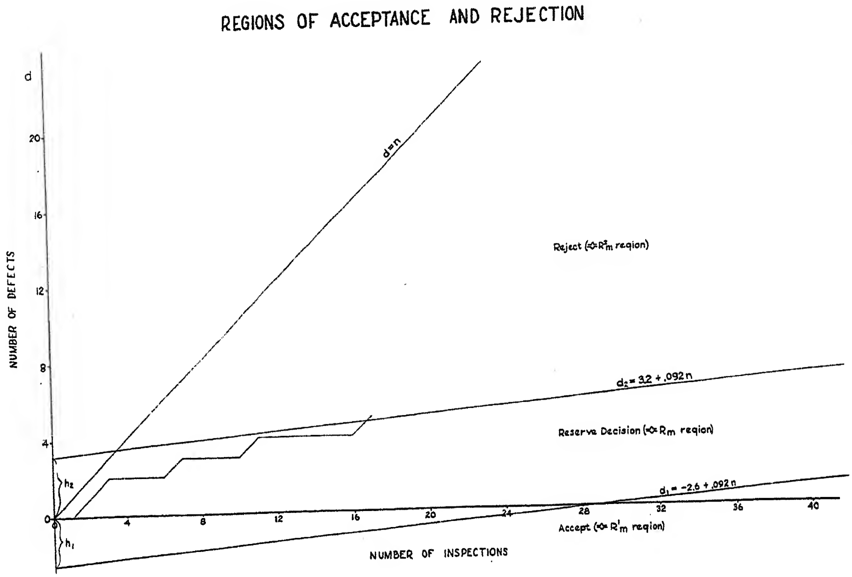
β is the risk of accepting a lot in which $p = p_2$.

$\alpha + \beta < 1$.

The operating characteristic or OC curve, as shown in Chart XIII XIII, indicates the relationship between p_1, p_2, α, and β. The ordinate L_p is the relative frequency of acceptance of a lot having p proportion of defectives under the particular sampling plan chosen. The ideal OC curve would be ⎺⎿ shaped. There would be some value p' ($p' = s$ as given by [13:156]) such that all lots having a smaller p would be accepted indubitably. Such a situation can be brought about if the inspection plan calls for sampling the entire lot.

It may be observed at this point that standard statistical procedures antedating the development of sequential analysis would endeavor to estimate from a sample the value of p in a lot, compare it with a standard proportion s, co-determinate, as shown later, with p_1 and p_2, and draw a conclusion as to acceptance or rejection. In standard statistics each of the two hypotheses can be tested independently. In sequential analysis a decision

CHART XIII XII



between the two is forced and must be accepted. Properly used, there is no systematic difference in outcomes, but it has been pointed out that the sequential analysis method will generally result in savings in the number of cases sampled in the neighborhood of 50 per cent. The question of the task placed upon the executive who decides upon the sampling scheme is also pertinent. If it is a sequential analysis scheme, he must choose, from a priori considerations, the crucial values p_1, p_2, α, and β, while if it is a standard scheme he must choose s and acceptable probability limits both for p below and for p above s.

The bounding lines d_1 and d_2 of Chart XIII XII are functions of h_1 and h_2, which are functions of p_1, p_2, α, and β, and of s which is a function of p_1 and p_2 only.

$$h_1 = \frac{\log\left(\frac{1-\alpha}{\beta}\right)}{\log\frac{p_2}{p_1}\left(\frac{1-p_1}{1-p_2}\right)} = \frac{b\ s}{\log\left(\frac{1-p_1}{1-p_2}\right)} \qquad [13:154]$$

$$h_2 = \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\log\frac{p_2}{p_1}\left(\frac{1-p_1}{1-p_2}\right)} = \frac{a\ s}{\log\left(\frac{1-p_1}{1-p_2}\right)} \qquad [13:155]$$

$$s = \frac{\log\left(\frac{1-p_1}{1-p_2}\right)}{\log\frac{p_2}{p_1}\left(\frac{1-p_1}{1-p_2}\right)} \qquad [13:156]$$

$$d_1 = -h_1 + sn \quad \text{Below which is acceptance region} \qquad [13:157]$$

$$d_2 = h_2 + sn \quad \text{Above which is rejection region.} \qquad [13:158]$$

The proof of these formulas is given in the revised report of the Statistical Research Group of the Applied Mathematics Panel (1945). Other important contributions to the early development of sequential analysis are Neyman and Pearson (1936), Hotelling (1941), Bernbaum (1942), Bartky (1943), Wald (1943), (1944 general, (1944) cumulative), (1945); Freeman (1944), Wald and Wolfowitz (1944), (1945).

The bounding lines of the chart can be plotted prior to the inspection of the first item. If upon inspection the first item is perfect, the point with coordinates (0,1) is plotted, while if defective, the point (1,0) is plotted. In either case, the plotted point lies in the region indicating further inspection as necessary, so a second observation is taken and the outcome plotted and

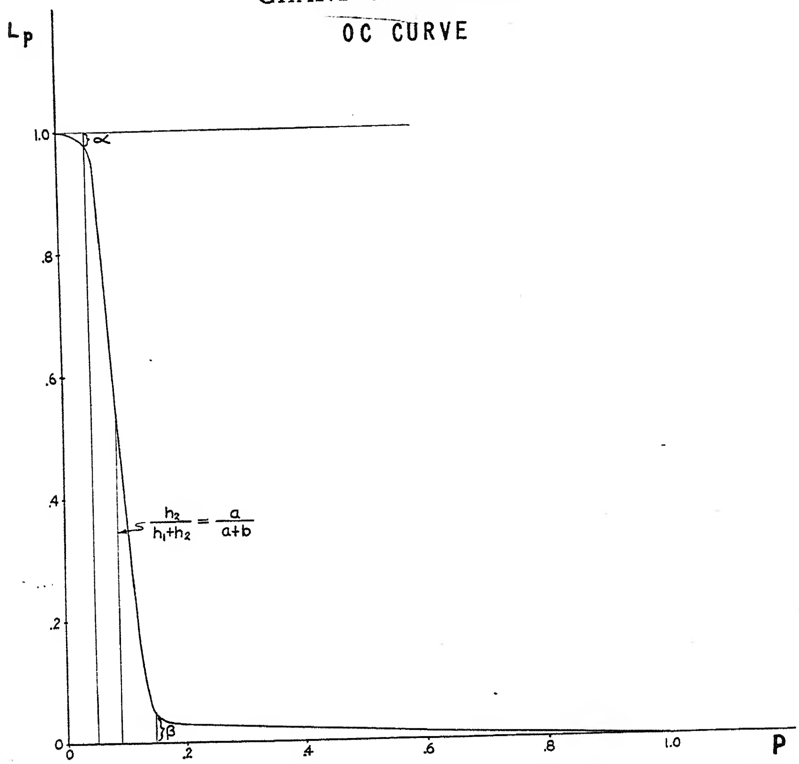
the process is continued until a plotted point is obtained in the region of rejection or in that of acceptance, thus terminating the inspection.

Alternative to this graphic procedure, one may draw up a table giving, for each successive value of n, a value of d above which lies rejection and a value of d below which lies acceptance.

The operating characteristic curve (the OC curve) of Chart XIII XIII is derivable from the

CHART XIII XIII

OC CURVE



properties of Wald's (1945 sequential) probability, or likelihood, ratio L.

$$L = \frac{P_2}{P_1} \quad \ldots\ldots\ldots \quad [13:159]$$

This is the ratio of the probability of the observed data arising if the second hypothesis, viz. $p = p_2$, is true to the probability of its arising under the alternate hypothesis, viz. $p=p_1$.

Two constant values of L, which we designate A and B, exist, which divide the m-dimensional space m_n ($m = 1,2,\ldots,n$) into three mutually exclusive and exhaustive parts, R_m^1, R_m^2, and R_m. Corresponding to the division of the m-dimensional space is a division as shown of the space in Chart XIII XII. If the sub-samples consist of one observation each, the m dimensions are the m observations, or scores, 0 or 1. R_m^1 is the region in the m-space wherein the $p=p_1$ hypothesis is tenable with a risk not greater than α. R_m^2 is the region wherein the alternative hypothesis $p=p_2$ is tenable with a risk not greater than β. R_m is the remaining region and herein neither of these hypotheses is tenable with the small risk assigned to it.

The sequential analysis is terminated at the smallest value n of m for which the sample point lies either in R_m^1 or R_m^2. If in R_m^1, we accept the first hypothesis and if in R_m^2, we accept the second hypothesis.

We may take the following as the values of A and β:

$$A = \frac{1-\beta}{\alpha}. \quad . \quad . \quad . \quad . [13:160]$$

$$B = \frac{\beta}{1-\alpha}. \quad . \quad . \quad . \quad . [13:161]$$

These are close approximations if α and β are small and the approximation is such as not to decrease the strength of the test, but perhaps increase very slightly the necessary number of observations. The logarithms of A and $1/B$ are needed so we write

$$a = \log A = \log \frac{1-\beta}{\alpha} \qquad \text{Note [13:170]} \qquad [13:162]$$

and

$$b = \log \frac{1}{B} = \log \frac{1-\alpha}{\beta} \qquad \text{Note [13:171]} \qquad [13:163]$$

It will be useful to note that

$$\frac{a}{a+b} = \frac{h_2}{h_1 + h_2} \quad \ldots\ldots \quad [13:164]$$

If, in Chart XIII XII, an ordinate is drawn at any point n and bounded by the lines $d = 0$ and $d = n$, it is divided into three parts. The ratio of the part lying in the reserve decision region to the sum of the parts in the other regions may be computed. This ratio obviously decreases as n increases, which is to say that, for practical purposes, if the sequential analysis is carried far enough, a point outside the reserve decision region will be attained, thus terminating the problem.

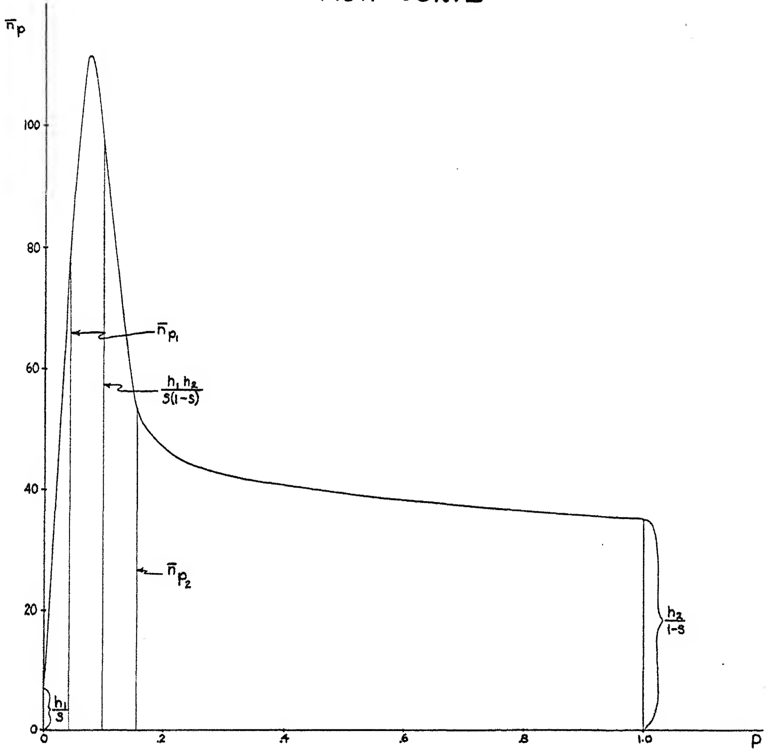
We can note that if, after a very large number of observations, a point in the reserve decision region is attained, the ratio of the number of defectives to the number of observations approaches the slope of the lines bounding this region, thus this slope, s as defined in [13:156], is also p', the standard value of p between p_1 and p_2, for which it is indifferent whether the lot is accepted or rejected. Note that s is a function of p_1 and p_2 only and not of the risks α and β.

Reference to Chart XIII XII shows that the value of n given by the intersection of $d = h_1 + sn$ and $d = 0$ gives the minimal number of observations which could lead to the acceptance of the lot. The value of n given by the intersection of $d = h_2 + s\,n$ and $d = n$ gives the minimal number of observations which could lead to the rejection of the lot. Ordinarily some number of

observations greater than either of these is necessary to reach a conclusion. The average sample size, $\bar{n}_p$, necessary to reach a decision, forms a distribution somewhat as indicated in Chart XIII XIV. We let the average sample number (ASN), required when $p = p_1$, be designated $\bar{n}_{p_1}$, and when $p = p_2$, we designate it $\bar{n}_{p_2}$. These may be the average sample numbers with which one is most concerned but, since the lot p is likely to lie between p_1 and p_2, a generally more informative ASN is $\bar{n}_s$. Certain ASN values are

CHART XIII XIV

ASN CURVE



approximately as follows:

$$\bar{n}_{p_1} = \frac{(1-\alpha)h_1 - \alpha\, h_2}{s - p_1} \quad . \; . \; . \; . \quad [13:165]$$

$$\bar{n}_{p_2} = \frac{(1-\beta)h_2 - \beta\, h_1}{p_2 - s} \quad . \; . \; . \; . \quad [13:166]$$

A good idea of the entire ASN curve can be gotten by plotting five points, $\bar{n}_{p_1}$, $\bar{n}_{p_2}$, and the following:

$$\bar{n}_0 = \frac{h_1}{s} \quad . \; . \; . \; . \; . \; . \; . \; . \quad [13:167]$$

$$\bar{n}_1 = \frac{h_2}{1 - s} \quad . \; . \; . \; . \; . \; . \quad [13:168]$$

$$\bar{n}_s = \frac{h_1\, h_2}{s(1 - s)} \quad . \; . \; . \; . \; . \; . \quad [13:169]$$

$\bar{n}_s$ is not the maximum point of the curve, but as indicated in Chart XIII XIV it lies near this maximum point.

The cost and time arrangements in connection with an inspection scheme should be sufficiently flexible to meet a situation in which the actual number of inspections required to reach a conclusion is 2.5 or 3 times as great as that indicated by the greater of $\bar{n}_{p_1}$ and $\bar{n}_{p_2}$.

The sequential analysis just discussed is that known as the "binomial case""because the scores, defective or non-defective, yields a binomial distribution. We here give a numerical illustration of this case and then discuss the normal (wherein the distribution of scores is normal), the Poisson, and other cases.

Illustrative numerical problem. A lot of 10,000 bolts is to be inspected to determine

whether it shall be accepted or rejected. The standards set by the accepting agency are:

p_1 = .05, i.e., if not over 5 per cent are defective with respect to the characteristic in question, the lot is acceptable.

α = .02, i.e., the risk of rejecting so good a lot is not greater than .02.

p_2 = .15, i.e., if over 15 per cent are defective, the lot is not acceptable.

β = .04, i.e., the risk of accepting so poor a lot is not greater than .04.

We compute by [13:154], [13:155], [13:156], [13:169], [13:157], [13:158].

$h_1 = 2.64$

$h_2 = 3.20$

$s = .09194$

$\bar{n}_s = 101$ 2 times 101 is judged not to be a prohibitive number of inspections.

$d_1 = 2.6 + .092\,n$

$d_2 = 3.2 + .092\,n$

As a result of inspection, the performance curve shown in Chart XIII XII was obtained, leading to rejection after but 17 bolts were inspected. Since 17 is considerably smaller than 51 ($= \bar{n}_{p_2}$), it is probable that the actual percentage of defectives is in excess of 15.

TABLE XIII J

n	d_1 ACCEPTANCE NUMBER	d NUMBER OF DEFECTIVES OBSERVED	d_2 REJECTION NUMBER
1	—	0	—
2	—	1	—
3	—	2	—
4	—	2	4
5	—	2	4
6	—	2	4
7	—	3	4
8	—	3	4
9	—	3	5
10	—	3	5
11	—	4	5
12	—	4	5
13	—	4	5
14	—	4	5
15	—	4	5
16	—	4	5
17	—	5	5
18	—		5
19	—		5
20	—		6
21	—		6
22	—		6
23	—		6
24	—		6
25	—		6
26	—		6
27	—		6
28	—		6
29	0		6
30	0		6
31	0		7
32	0		7
33	0		7
34	0		7
35	0		7
36	0		7
37	0		7
38	0		7
39	0		7
40	1		7
41	1		8
42	1		8
43	1		8
44	1		8
45	1		8
.	.		.
.	.		.
.	.		.
202	15		22

The tabular treatment of Table XIII J is equivalent to this graphic procedure. Columns n, d_1, and d_2 are drawn up prior to the first inspection. The recorded d_2 values are the integral values next higher than fractional values ordinarily given by [13:158] and the recorded d_1 values are the integral values next smaller than fractional values given by [13:157]. The entries in column d are recorded as 1, 2, 3, etc., inspections are made. Since, after 17 inspections, a value in the d column is attained as great as the value in the d_2 column, the lot is rejected. If the lot had been a much better lot, a value would have been attained, after a certain number of inspections, as small as the value in the d_1 column and the lot would have been accepted.

The normal case (involving a one-sided alternative test of the mean): If the score, X, attaching to an observation is a normally distributed score with known standard deviation, σ, (and variance V) a precise sequential analysis test of the difference between the two means is available. We have

M_1 the smaller mean

M_2 the larger mean

α the risk of believing a lot to have a mean of M_2 when it in fact has a mean of M_1

β the risk of believing a lot to have a mean of M_1 when it in fact has a mean of M_2

We here desire, for purposes of acceptance, lots whose means exceed M_2 and this is a one-sided alternative. Had we desired lots whose means neither exceeded nor fell short of M_2 by a certain amount we would have had a two-sided alternative, a readily soluble problem (SRG Report 255, Sec. 5, 1945).

Let a and b be the following natural logarithms:

$$a = \ln\frac{1-\beta}{\alpha} \quad \text{Note [13:162]} \qquad [13:170]$$

$$b = \ln\frac{1-\alpha}{\beta} \quad \text{Note [13:163]} \qquad [13:171]$$

We have for the indifferent mean and also for the slope of the acceptance and rejection boundary lines

$$s = \frac{M_1 + M_2}{2} \quad \ldots\ldots\ldots [13:172]$$

The intercepts of these boundary lines are h_2 and $-h_1$.

$$h_1 = \frac{b\ V}{M_2 - M_1} \quad \ldots\ldots\ldots [13:173]$$

$$h_2 = \frac{a\ V}{M_2 - M_1} \quad \ldots\ldots\ldots [13:174]$$

The acceptance and rejection boundary lines are given by [13:157] and [13:158]. The operating characteristic curve (a left-right reflection of a Chart XIII XIII curve, with abscissa labeled M instead of p) is

$$L_M = \frac{e^{t_1} - 1}{e^{t_2} - 1} \quad \ldots\ldots\ldots [13:175]$$

in which

$$t_1 = 2\ (s-M)h_1 \quad \ldots\ldots [13:176]$$

$$t_2 = 2\ (s-M)\ \ (h_1+h_2) \ldots. [13:177]$$

L_M is the relative frequency of accepting a lot whose mean quality is M. L_M is of indeterminate form when $M = s$, but it can be evaluated. We

then have

$$L_s = \frac{h_1}{h_1 + h_2} \quad \ldots\ldots\ldots\ldots [13:178]$$

Average sample numbers are readily obtained.

$$\bar{n}_M = \frac{L_M\ (h_1 + h_2) - h_1}{M - s} \quad \ldots\ldots [13:179]$$

$$\bar{n}_s = \frac{h_1 h_2}{V} \ldots\ldots\ldots\ldots \quad [13:180]$$

$$\bar{n}_{M_1} = \frac{2[(1-\alpha)\ b - \alpha\ a]\ V}{(M_2 - M_1)^2} \qquad [13:181]$$

$$\bar{n}_{M_2} = \frac{2[(1-\beta)\ a - \beta\ b]\ V}{(M_2 - M_1)^2} \qquad [13:182]$$

We need criteria for the acceptance of $M = M_1$ and of $M = M_2$. Since we have a normal distribution the ratio of the likelihood that the sample in question would arise under the hypothesis $M = M_2$ to its likelihood under the alternative hypothesis $M = M_1$ is simply,

$$L = \frac{\exp -\frac{1}{2V} \Sigma(X - M_2)^2}{\exp -\frac{1}{2V} \Sigma(X - M_1)^2} \quad \ldots\ldots [13:183]$$

Taking natural logarithms, which does not disturb the relationships [13:162] and [13:163], we can readily obtain the following criteria:

$$\Sigma X - ns \geq \frac{V\ a}{M_2 - M_1} \quad \text{Accept } M = M_2 \qquad [13:184]$$

$$\Sigma X - ns \leqq \frac{V\,a}{M_2 - M_1} \qquad \text{Accept } M = M_1 \qquad [13:185]$$

$$\frac{-V\,b}{M_2 - M_1} < (\Sigma X - ns) < \frac{V\,a}{M_2 - M_1} \qquad \text{Reserve decision, i.e., continue testing} \qquad [13:186]$$

The actual solution may be carried out graphically by means of such a chart as XIII XII or by tabular methods, recording, for successive n's, the values of ΣX which lead to the acceptance of $M = M_2$ and again of $M = M_1$.

Further applications of sequential analysis: The method applies to many situations of which the following will be noted:

Double dichotomies: (SRG 255, Sec. 3) revised, 1945) Herein the problem is that of choosing between two processes, each of which leads to one of two results. An observation (random) of a first process case is compared with an observation (random) of a second process case, with one of three outcomes, (*a*) first process superior, (*b*) first and second processes equal, or (*c*) first process inferior. This is continued until the sequential analysis test shows one of the processes to be superior, within the risk limits set.

Test of variability: (SRG 255, Sec. 6 revised, 1945) The question here is to decide whether the standard deviation of a submitted lot differs (within some small assigned risk) from some standard, or acceptable, standard deviation. The solution is dependent upon a normal distribution.

Test in the case of a Poisson distribution: (SRG 255, Sec. 7, 1945) The chief hazard here lies in the assumption of a Poisson distribution. The experimental determination that a Poisson distribution is present is in general far more difficult than the sequential analysis procedure that follows therefrom.

CHAPTER XIV

MATHEMATICS, THE MENTOR OF STATISTICAL INGENUITY

SECTION 1. INGENUITY IN RESEARCH

Cleverness in laboratory technique, in utilizing unique properties of materials, and in measuring differences rather than gross magnitudes has resulted in many of the great discoveries of modern physical, biological, and even social science. Another field for ingenuity has been relatively neglected. It has been a tacit assumption that the elementary associative and distributive laws of mathematics hold. Generally this has proved serviceable and where it has not the disproof of the assumption has frequently been the most direct way of discovering some important truth. There are innumerable relationships between and within mathematical functions and unique to the particular functions employed, which have not been exploited in connection with phenomena. It seems that it has not been customary to look to mathematics for the key to material and biological behavior. The writer believes that mathematical relation-

ships are never the exact key and that living behavior is always more complex and always conditioned by more factors than are included in a mathematical formulation, but nevertheless such formulation may be an approximate key and one of the richest sources for suggestions as to life behavior.

A striking illustration of this principle is to be found in the properties of Latin and Graeco-Latin squares and in the design of experiments as developed by R. A. Fisher (1937) and his school. With plot designs consisting of rows and columns and equal numbers of observations in each cell, there is zero correlation between the frequencies in the rows and the frequencies in the columns. This is a property discovered through the study of mathematics. Agricultural experimentalists have been shrewd enough to incorporate this principle in their experimental designs and they have obtained an economy of procedure and a certainty of conviction not previously possible. It is in this sense that mathematics is to be thought the mentor of statistical ingenuity. The use of the properties of the normal distribution, the principles of sequential analysis, the merits of independent variables as availed of in the analysis of variance and in multivariate analysis into principal components, are a few additional illustrations of illuminating transfer from mathematics to living phenomena.

There are innumerable other mathematical properties which have never been availed of in experimental design and procedure. In the following sections are given a number of mathematical functions having unique properties which may be suggestive of phenomenological relationships. This chapter is intended as a background—though

all too brief—for the experimentalist so that, if his data do show some of the symptoms of say a particular mathematical function, he can then seek out all the invariant properties of the function and return to his data with the hypothesis that it has parallel invariant properties. Mathematics should be more than a cultivator that is dragged behind a lead horse,—it should be the horse itself, and frequently a spirited one, that chooses the ground that is to be cultivated.

SECTION 2. MATRICES AND DETERMINANTS

A few facts about matrices and determinants will assist in understanding many theoretical as well as applied problems in statistics. A matrix is a rectangular array of terms, called elements. It does not have a quantitative value. A matrix may be designated by a single letter (usually script capital $\mathcal{A}$, $\mathcal{B}$, etc.), by a complete recording of all its elements, by a partial recording with dots to indicate obvious non recorded elements, by a designation of the general element a_{ij}, or thus— $\|a_{ij}\|$. Also in the case of a square matrix, by giving the terms in the *principal diagonal*, thus $\|a_{11}, a_{22}, \ldots, a_{kk}\|$

$$\mathcal{A} = \|a_{11}, a_{22}, \ldots, a_{kk}\| = \begin{Vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{Vmatrix} \qquad [14:01]$$

The double rules are used to distinguish between

a matrix and a determinant.

A matrix having k rows and ℓ columns is of order $k\ell$.

Two matrices are equal only if each element of the one is equal to an element in corresponding position in the other.

Matrix $\mathfrak{a}'$ derived from matrix $\mathfrak{a}$ by interchanging rows and columns is called its *transpose*. E.g.

$$\mathfrak{a} = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{Vmatrix}$$

and [14:02]

$$\mathfrak{a}' = \begin{Vmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{Vmatrix}$$

The *adjoint of a square* matrix $\mathfrak{a}$ is the matrix whose elements are the cofactors (defined later) of the elements of $\mathfrak{a}'$.

A square matrix is *symmetric* if and only if it is equal to its transpose. Thus an element a_{ij} must equal its *conjugate* a_{ji}.

The addition of matrices: If three matrices $\mathfrak{a}$, $\mathfrak{B}$, and $\mathfrak{C}$, having elements a_{ij}, b_{ij}, and c_{ij} are such that $\mathfrak{C}$ is the *sum of* $\mathfrak{a}$ *and* $\mathfrak{B}$,

$$\mathfrak{C} = \mathfrak{a} + \mathfrak{B}$$

then $c_{ij} = a_{ij} + b_{ij}$, these elements being added in the ordinary algebraic sense.

Matrix multiplication by a scalar: If the matrix $\mathfrak{a}$ is *multiplied by a scalar* (a real number, as distinguished from a vector) b, then every element in $\mathfrak{a}$ is multiplied by b, thus, e.g.

$$\mathfrak{a} = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{Vmatrix} \text{ and } b\,\mathfrak{a} = \mathfrak{a}\,b = \begin{Vmatrix} ba_{11} & ba_{12} & ba_{13} \\ ba_{21} & ba_{22} & ba_{23} \end{Vmatrix} \quad [14:03]$$

The multiplication of matrices: With matrix theory meanings attached to the terms pre- and post-multiplication, let the matrix $\mathbb{C}$ be equal to matrix $\mathbb{A}$ post-multiplied by the matrix $\mathbb{B}$. The element in ith row and kth column of $\mathbb{C}$ is equal to the sum of the products of the successive elements in the ith row of $\mathbb{A}$ with the successive elements in the jth column. E.g., let

$$\mathbb{A} = \begin{Vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{Vmatrix}$$

and

$$\mathbb{B} = \begin{Vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{Vmatrix} \qquad [14:04]$$

then

$$\mathbb{C} = \mathbb{A}\,\mathbb{B} = \begin{Vmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} \end{Vmatrix}$$

$\mathbb{A}$ is *pre-multiplied* by $\mathbb{B}$, or $\mathbb{B}$ is *post-multiplied* by $\mathbb{A}$. The multiplication operation is not commutative, for in general $\mathbb{A}\mathbb{B} \neq \mathbb{B}\mathbb{A}$.

The transpose of the product of any number of matrices is equal to the product of their transposes taken in reverse order. E.g., $(ABC)' = C'B'A'$.

The identity matrix $\mathbb{I}$ is the following square matrix:

$$\mathbb{I} = \begin{Vmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 0 & 0 & . & . & . & 1 \end{Vmatrix} \ldots\ldots [14:05]$$

It is such that $\mathcal{A}\mathcal{I} = \mathcal{I}\mathcal{A} = \mathcal{A}$.

For a square matrix $\mathcal{A}$, with elements a_{ij}, let A = the determinant having the same elements, a_{ij}, as matrix $\mathcal{A}$, and let A_{ij} be the cofactors of the elements a_{ij}. Then the *inverse* of $\mathcal{A}$ is $\mathcal{A}^{-1}$ and is given by

$$\mathcal{A}^{-1} = \frac{1}{A}\begin{Vmatrix} A_{11} & A_{21} & \cdots & A_{k1} \\ A_{12} & A_{22} & \cdots & A_{k2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ A_{1k} & A_{2k} & \cdots & A_{kk} \end{Vmatrix} = \frac{1}{A} \times (\textit{adjoint of } \mathcal{A}) \qquad [14{:}06]$$

Note that the element that is at the intersection of the *i*th row and *j*th column in the adjoint is not the cofactor of a_{ij}, but of a_{ji}. A matrix and its inverse are such that

$$\mathcal{A}\mathcal{A}^{-1} = \mathcal{A}^{-1}\mathcal{A} = \mathcal{I} \quad . \; . \; . \; . \; . \; . \; . \quad [14{:}07]$$

The multiplication of matrices is associative. E.g., $\mathcal{A}\,\mathcal{B}\,\mathcal{C} = \mathcal{A}(\mathcal{B}\,\mathcal{C}) = (\mathcal{A}\,\mathcal{B})\mathcal{C}$, the order of the letters being the same in all instances.

The inverse of any product of matrices is the product of their inverses in reverse order.

The multiplication of matrices is distributive. E.g.

$$\mathcal{A}(\mathcal{B}+\mathcal{C}) = \mathcal{A}\,\mathcal{B} + \mathcal{A}\,\mathcal{C} \quad . \; . \; . \; . \; . \; . \; . \; . \quad [14{:}08]$$

A diagonal matrix has non-zero elements in the diagonal only. If these elements are the same throughout the matrix is a *scalar matrix*, it having the property that pre-multiplication by it and post-multiplication by it are the same, and further, if the common element in the diagonal terms of the scalar matrix $\mathcal{G}$

$$\mathcal{G} = \begin{Vmatrix} g & 0 & \cdots & 0 \\ 0 & g & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & g \end{Vmatrix} \qquad [14:09]$$

is g then $\mathcal{A}\mathcal{G} = \mathcal{G}\mathcal{A} = \mathcal{A}g = g\mathcal{A} =$ a matrix having elements (ga_{ij}).

In an equation of matrices a multiplying factor may be moved to the other member of the equation by substituting its inverse, while maintaining its same relative position. E.g.

$$\left.\begin{aligned} \mathcal{A}\,\mathcal{B}\,\mathcal{C} &= \mathcal{D} \\ \mathcal{B}\,\mathcal{C} &= \mathcal{A}^{-1}\mathcal{D} \\ \mathcal{C} &= \mathcal{B}^{-1}\mathcal{A}^{-1}\mathcal{D} \end{aligned}\right\} \text{also} \left\{\begin{aligned} \mathcal{A}\,\mathcal{B}\,\mathcal{C} &= \mathcal{D} \\ \mathcal{A} &= \mathcal{D}(\mathcal{B}\,\mathcal{C})^{-1} \\ \mathcal{A} &= \mathcal{D}\,\mathcal{C}^{-1}\mathcal{B}^{-1} \end{aligned}\right\}$$

$$\text{also} \quad \left\{\begin{aligned} \mathcal{A}\,\mathcal{B}\,\mathcal{C} &= \mathcal{D} \\ \mathcal{B}\,\mathcal{C} &= \mathcal{A}^{-1}\mathcal{D} \\ \mathcal{B} &= \mathcal{A}^{-1}\mathcal{D}\,\mathcal{C}^{-1} \end{aligned}\right. \qquad [14:10]$$

Determinants: A matrix, including a square matrix, is just an arrangement of elements, but if the elements in the square matrix $\mathcal{A}$, equal to $\|a_{11}\, a_{22} \ldots a_{kk}\|$, are evaluated as *the determinant A*, (also frequently designated Δ)

$$A = |a_{ij}| = |a_{11}a_{22} \ldots a_{kk}| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix} \qquad [14:11]$$

we have an ordinary algebraic polynomial.

A determinant of k rows (and k columns) is one of the kth order.

The $k!$ terms in the polynomial A cover all possible combinations yielding terms having one element from each row and each column. *The sign factor* that attaches to each term can be determined by ascertaining whether the number of inversions in the first subscripts plus the number in the second subscripts is even or odd. If even the sign is positive, and if odd it is negative. E.g., consider the following term in a fifth order determinant: $a_{15}a_{23}a_{32}a_{41}a_{54}$. The first subscripts, 1, 2, 3, 4, 5, show no inversions and the second subscripts 5, 3, 2, 1, 4, show seven (the 5 being to the left of four smaller numbers, the 3 to the left of two, and the 2 to the left of one). Accordingly, $-a_{15}a_{23}a_{32}a_{41}a_{34}$ is one of the 5! terms in the expanded fifth order determinant.

a_{11} is *the leading element* of the determinant A.

$a_{11}a_{22} \ldots a_{kk}$ is *the leading term.*

The a_{11} to a_{kk} is *the principal diagonal,* and the a_{k1} to a_{1k} is *the secondary diagonal.*

Interchanging two rows, or two columns, changes the sign of the determinant.

If the ith row and the jth column of A are crossed out, the remaining determinant is called *a first minor* and frequently designated Δ_{ij}. A first minor of the type Δ_{ii} is *a principal first minor.*

A positive or negative sign attaches to the i, j position. It is given by $(-1)^{i+j}$. When this sign is attached to the minor we have *a cofactor,* which may be designated A_{ij}.

$$A_{ij} = (-1)^{i+j}\,\Delta_{ij} \quad . \; . \; . \; . \qquad [14:12]$$

A determinant may be expanded in terms of the elements of any row, or column, and their cofactors. E.g., in terms of the elements of the third row:

$$A = a_{31}A_{31} + a_{32}A_{32} + \dots + a_{3k}A_{3k}, \text{ or}$$
$$A = a_{31}\Delta_{31} - a_{32}\Delta_{32} + \dots + (-1)^{3+k}a_{3k}\Delta_{3k} \qquad [14:13]$$

The terms symmetric, positive definite, Gramian apply to determinants as well as to square matrices.

A positive definite determinant is one in which all the principal minors > 0.

A gramian determinant is a symmetric positive definite determinant. This is the characteristic type of major determinant in multiple correlation and regression.

The rank of a non-vanishing determinant is equal to the order of the determinant.

The rank of a vanishing determinant is equal to the rank of its largest non-zero minor. The rank of a matrix is equal to the rank of the determinant of largest rank that can be made from its rows and columns.

If the elements of any row, or column, of a determinant are multiplied by a constant, *the determinant is multiplied by this constant.*

If the elements of any row (or column) multiplied by a constant yield the elements of any other row (or column), then the determinant is equal to zero.

If the addition of α_{ij} to an element, a_{ij}, of the non-vanishing determinant A, yields a determinant which equals zero, then α_{ij} is a factor of A, and a_{ij} does not appear in the expanded and and reduced form of A. If $\alpha_{1a}, \alpha_{2b}, \dots, \alpha_{kg}$ are k such factors, where k is the order of A, $A = \alpha_{1a}\alpha_{2b} \dots \alpha_{kg}$ (a sign factor). If the num-

ber of inversions in the order $a, b, \ldots, g$ is even, the sign is positive, and if odd it is negative. In particular should the factors be $\alpha_{11}, \alpha_{22}, \ldots \alpha_{kk}$, then

$$A = \alpha_{11}\ \alpha_{22} \cdots \alpha_{kk} \qquad [14:14]$$

DETERMINANTAL SOLUTION OF SIMULTANEOUS EQUATIONS

Given k simultaneous linear equations, so written that the constant terms are to the right of the = sign, thus:

$$\begin{array}{l}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{kk}x_k = \alpha_{01} = \alpha_{10} \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{ak}x_k = \alpha_{02} = \alpha_{20} \\
\cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\
\cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\
\cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\
a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k = \alpha_{0k} = \alpha_{k0}
\end{array} \qquad [14:15]$$

Write down the major determinant

$$\Delta = A = \begin{vmatrix}
a_{11} & a_{12} & \cdots & a_{1k} & a_{10} \\
a_{21} & a_{22} & \cdots & a_{2k} & a_{20} \\
\cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & & \cdot & \cdot \\
a_{k1} & a_{k2} & \cdots & a_{kk} & a_{k0} \\
a_{01} & a_{02} & \cdots & a_{0k} & a_{00}
\end{vmatrix} \qquad [14:16]$$

in which the value of a_{00} is immaterial in connection with the solution for the x's, as it disappears in all the equations [14:17]. The values of the x's are given by

$$x_1 = \frac{\Delta_{01}}{\Delta_{00}} = \frac{-A_{01}}{A_{00}}$$

$$x_2 = \frac{-\Delta_{02}}{\Delta_{00}} = \frac{-A_{02}}{A_{00}} \qquad [14:17]$$

$$x_3 = \frac{\Delta_{03}}{\Delta_{00}} = \frac{-A_{03}}{A_{00}}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$x_k = \frac{(-1)^{k+1}\Delta_{0k}}{\Delta_{00}} = \frac{-A_{0k}}{A_{00}}$$

SECTION 3. THE POINT BINOMIAL

The moments of $(p+q)^n$ about its mean are readily found by use of a reduction formula due to Romanovsky* (1923). We have

Moments of a binomial series

$\mu_0 = 1$ . . . . [14:19]

$\mu_1 = 0$ . . . . [14:20]

$V = \mu_2 = npq$ . . . . See [9:02] . . . . . . [14:21]

$\mu_3 = npq(q-p)$. . See [9:03] . . . . . . [14:22]

* Romanovsky shows that

$$\mu_{s+1} = pq\left(\frac{d\mu_s}{dp} + n\,s\,\mu_{s-1}\right) \qquad [14:18]$$

The repeated use of this gives μ_2, μ_3, μ_4, etc.

$$\mu_4 = npq[1+3(n-2)pq] \quad \text{. . See [9:04] . . . .} \quad [14:23]$$

$$\mu_5 = npq(q-p)[1+pq(-12+10n)] \quad \text{. . . . . .} \quad [14:24]$$

$$\mu_6 = npq\{1+5pq[-6+5n+pq(24-26n+3n^2)]\} \quad [14:25]$$

SECTION 4. THE POISSON DISTRIBUTION

The Poisson series is obtained from the binomial by letting p, the probability of the event occurring, be indefinitely small, but n so large that pn is finite. Poisson's exponential limit of the binomial $(p+q)^n$ (see Yule and Kendall, 1937, pp. 187-191) is a distribution such that the probability of a variate (x = 0, 1, 2 ...) taking the value of x is $M^x e^{-x}/x!$, or

$$e^{-M}\left(1+M+\frac{M^2}{2!}+\frac{M^3}{3!}+\ \ldots\ \right) \quad \text{Poisson distribution} \quad [14:26]$$

e^{-M} is the relative frequency of X = 0; $e^{-M}M$ the relative frequency of X = 1; $e^{-M}M^2/2!$ the relative frequency of X = 2; etc. The first four moments of this distribution about the point X=0 are

$$\mu_1' = M = pn \quad \text{. . . . .} \quad [14:27]$$

$$\mu_2' = M(M+1) \quad \text{. . . . .} \quad [14:28]$$

$$\mu_3' = M(M^2+3M+1) \quad \text{. . . . .} \quad [14:29]$$

$$\mu_4' = M(M^3+6M^2+7M+1) \quad \text{. . . . .} \quad [14:30]$$

The moments from M are

Moments of a Poisson series

$$\mu_1 = 0 \quad \text{. . . . .} \quad [14:31]$$

$$V = \mu_2 = M \quad \text{. . . . .} \quad [14:32]$$

$$\mu_3 = M \quad \text{. . . . .} \quad [14:33]$$

$$\mu_4 = M(1+3M) \quad \ldots\ldots \quad [14:34]$$

$$\mu_5 = M(1+10M) \quad \ldots\ldots \quad [14:35]$$

$$\mu_6 = M(1+25M+15M^2) \quad \ldots\ldots \quad [14:36]$$

The moments are determinable from a knowledge of the mean and without knowledge of p and n separately. Thus for the Poisson, M is a *sufficient statistic*.

Molina (1942) has tabled individual terms and cumulated sums for M = 0 to .01, by intervals of .001; for M = .01 to 15 by intervals of .01; for M = 15 to 100 by intervals of 1. It seems probable that all ordinary needs can be served by this table.

Pearson (1914) has tabled the successive terms of the Poisson distribution, to six decimal places, for M from .1 to 15, by tenths. All the cumulants of a Poisson distribution are equal (and equal to M), a fact not true for any other distribution. If $x_3 = x_1 + x_2$, in which x_1 is distributed in the Poisson manner with parameter M, and x_2, independent of x_1, is distributed in the Poisson manner with parameter M', then x_3 is distributed in the Poisson manner with parameter $(M+M')$.

SECTION 5. THE HYPERGEOMETRIC SERIES

This series can be explained by reference to an example. Let there be an urn containing n balls, of which qn are black and pn are white. Let N balls be drawn at a time. Let X equal the number of white balls in such a sample of N. This number will vary from sample to sample from 0 to N, or to pn, whichever is the smaller.

When N are drawn the probability that the first of the N is not-white is qn/n; that this having come to pass the second is not-white it is $(qn-1)/(n-1)$; that, the preceding having come to pass, the third is not-white it is $(qn-2)$

$(N-2)$; etc. to $(qn-N+1)/(n-N+1)$, the probability that the preceding having been not-white the Nth is also not-white. The product of all these is the probability that when N are drawn none will be white. This gives the proportionate probability given in the first row of the accompanying table. The other rows are obtained by a very similar procedure.

X = no. of white balls	Proportionate frequency of occurrence	
0	$\dfrac{qn(qn-1)(qn-2)\ldots(qn-N+1)}{n(n-1)(n-2)\ldots(n-N+1)}$	
1	$\dfrac{pn(qn)(qn-1)\ldots(qn-N+2)}{n(n-1)(n-2)\ldots(n-N+1)} \times C_1^N$	
2	$\dfrac{pn(pn-1)(qn)\ldots(qn-N+3)}{n(n-1)(n-2)\ldots(n-N+1)} \times C_2^N$	
$\vdots$	$\vdots$	[14:37]
X	$\dfrac{(pn)!\ (qn)!\ N!\ (n-N)!}{n!\ (pn-X)!\ (qn-N+X)!\ (N-X)!\ X!}$	
$\vdots$	$\vdots$	
N	$\dfrac{pn(pn-1)(pn-2)\ldots(pn-N+1)}{n(n-1)(n-2)\ldots(n-N+1)}$	

This series of frequencies is a hypergeometric series. When n is infinitely large with reference to N it becomes a binomial series.

Pearson (1924) has provided formulas for the first four moments as herewith:

Moments of a hypergeometric series

$$M = pN \quad \ldots \quad [14:38]$$

$$\mu_0 = 1 \quad \ldots \quad [14:39]$$

$$\mu_1 = 0 \quad \ldots \quad [14:40]$$

$$\mu_2 = \frac{Npq(n-N)}{(n-1)} \quad \ldots \quad [14:41]$$

$$\mu_3 = \frac{Npq(q-p)(n-N)(n-2N)}{(n-1)(n-2)} \quad \ldots \quad [14:42]$$

$$\mu_4 = \frac{Npq(n-N)\{n(n+1)-6N(n-N)+3pq[n^2(N-2)-nN^2+6N(n-N)]\}}{(n-1)(n-2)(n-3)} \quad [14:43]$$

An important statistic in connection with a probability function such as the binomial, the Poisson, or the hypergeometric distribution, is the sum of the frequencies up to a designated point. This matter is adequately covered by tables in connection with the Poisson distribution. Camp (1924 and 1925) gives methods for obtaining this desired sum for both the binomial and hypergeometric distributions. They would seem to be adequate for most practical needs, though not of the highest precision nor of the simplicity of computation which one would desire.

The obtaining of this sum for the still more useful normal, t (Student's t), χ^2, Pearson Type III, and variance ratio distributions is, in all cases, possible, as explained elsewhere in this text.

SECTION 6. FACTORIALS AND THE GAMMA FUNCTION

The gamma function of a number $x+1$ is defined by

$$\Gamma(x+1) = \int_0^\infty e^{-y}\, y^x\, dy \quad \text{The gamma function} \quad [14:44]$$

It is such that

$$\Gamma(x+1) = x\,\Gamma x \; . \; . \; . \; . \; . \; \Gamma \text{ reduction formula} \quad [14:45]$$

This holds for all values of x, integral, fractional, positive or negative. The term factorial is commonly applied to positive integers, in which case

$$x! = x\,(x-1)! \quad \text{Factorial reduction formula} \quad [14:46]$$

When x is a positive integer

$$\Gamma(x+1) = x! \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [14:47]$$

It is frequently serviceable to generalize the concept factorial so that [14:46] and [14:47] hold when x is not a positive integer. When this is done we note that by repeated applications of [14:45] the factorial of any number can be expressed as the product of certain numbers times the factorial of a number x', wherein $0 \leqq x' \leqq 1$. For example

$$2.5! = 2.5 \times 1.5! = 2.5 \times 1.5 \times .5!$$

Another example:

$$(-2.5)! = \frac{(-1.5)!}{-1.5} = \frac{(-.5)!}{(-1.5)\times(-.5)} = \frac{.5!}{(-1.5)\times(-.5)\times.5}$$

Thus, except for the labor involved, a table of $x!$, for values from .00 to 1.00, or of Γx from 1.00 to 2.00, is sufficient to yield, with operations of multiplication and division, all factorials and all gamma functions. We provide herewith a brief table of this sort.

TABLE XIV A

BASIC FACTORIALS AND GAMMA FUNCTIONS

x	$x!$, or $\Gamma(x+1)$	x	$x!$, or $\Gamma(x+1)$	x	$x!$, or $\Gamma(x+1)$	x	$x!$, or $\Gamma(x+1)$
−.01	1.00587						
.00	1.00000	.25	.90640	.50	.88623	.75	.91906
.01	.99433	.26	.90440	.51	.88659	.76	.92137
.02	.98884	.27	.90250	.52	.88704	.77	.92376
.03	.98355	.28	.90072	.53	.88757	.78	.92623
.04	.97844	.29	.89904	.54	.88818	.79	.92877
.05	.97350	.30	.89747	.55	.88887	.80	.93138
.06	.96874	.31	.89600	.56	.88964	.81	.93408
.07	.96415	.32	.89464	.57	.89049	.82	.93685
.08	.95973	.33	.89338	.58	.89142	.83	.93969
.09	.95546	.34	.89222	.59	.89243	.84	.94261
.10	.95135	.35	.89115	.60	.89352	.85	.94561
.11	.94740	.36	.89018	.61	.89468	.86	.94869
.12	.94359	.37	.88931	.62	.89592	.87	.95184
.13	.93993	.38	.88854	.63	.89724	.88	.95507
.14	.93642	.39	.88785	.64	.89864	.89	.95838
.15	.93304	.40	.88726	.65	.90012	.90	.96176
.16	.92980	.41	.88676	.66	.90167	.91	.96523
.17	.92670	.42	.88636	.67	.90330	.92	.96877
.18	.92373	.43	.88604	.68	.90500	.93	.97240
.19	.92089	.44	.88581	.69	.90678	.94	.97610
.20	.91817	.45	.88566	.70	.90864	.95	.97988
.21	.91558	.46	.88560	.71	.91057	.96	.98374
.22	.91311	.47	.88563	.72	.91258	.97	.98766
.23	.91075	.48	.88575	.73	.91467	.98	.99171
.24	.90852	.49	.88595	.74	.91683	.99	.99581
						1.00	1.00000
						1.01	1.00427
E^{II}	.00002		.00001		.00001		.00001
E^{III}	.00000		.00000		.00000		.00000

The E_2 values recorded at the bottom of the columns give the maximal 2-point, or linear, interpolation error as determined for entries half way down the column. Similarly the maximal 3-point, or quadric, interpolation error is given.

For gammas and factorials of small numbers the reduction formulas [14:45] and [14:46] may first be employed and then the necessary values found in Table XIV A. For large numbers, $x > 4$, Stirling's formula

$$\ln(x!) = (x+\frac{1}{2})\ln x - x + \frac{1}{2}\ln(2\pi) + \frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \frac{1}{1680x^7} + \dots \qquad [14:48]$$

and Forsyth's

$$\Gamma(x+1) = \sqrt{2\pi}\left(\frac{\sqrt{\frac{1}{6} + x + x^2}}{e}\right)^{x+\frac{1}{2}} \qquad [14:49]$$

are highly reliable. If n is large the error in Forsyth's formula is less than $1/(240x^3)$ of the whole. (See Pearson 1895 and 1901). Even for $x = 1.5$ the error is only in the neighborhood of 1 per cent. (See Kelley 1923 p. 136).

A function frequently needed in curve fitting is $\Gamma(x+1)/x^x e^{-x}$. The logarithm of this to the base 10 as given by Pearson (1934) is

$$\log\frac{\Gamma(x+1)}{x^x e^{-x}} = .3990899 + \frac{1}{2}\log x + .080929 \sin\frac{25^\circ 623}{x} \qquad [14:50]$$

The incomplete gamma function: If the upper limit of the integral [14:44] is changed from ∞ to y, we have an "incomplete gamma function," which has been tabled by Pearson. This is important in that this is the probability integral of a Pearson Type III distribution,—the equation being written with the origin at the mode. The Type III distribution has been found to be an excellent approximation to much biological and physical phenomena. One illustration is the distribution of yearly maximum flow of water in a river.

SECTION 7. THE NUMERICAL SOLUTION OF HIGHER DEGREE PARABOLIC AND OF EXPONENTIAL EQUATIONS

So many procedures have been employed that the student may expect to find in the mathematical literature (see Scarborough, 1930, Ch. IX and Whittaker and Robinson, 1924, Ch. VI) a method adapted to his particular problem. Only two are given here, the first being serviceable in the case of a rapidly convergent power series and the second when the function can be expressed as the sum or the product of two functions, each sufficiently simple to permit of rapid plotting.

If

$$y = a + bx + cx^2 + dx^3 + ex^4 + \text{negligible terms} \quad [14:51]$$

We let x_1 be a first estimate of the correct value. Substitute x_1 in all terms of [14:51] beyond the x term and solve for x, calling the value obtained x_2, a second estimate. Substitute x_2 in terms beyond the x term, solve, and call the value obtained x_3, a third estimate. Continue the iterative process until two successive values, say x_3 and x_4, differ by a small, but not quite negligible amount. Compute y_3 and y_4:

$$y_3 = a + bx_3 + cx_3^2 + dx_3^3 + ex_3^4 \; . \; . \; . \; . \quad [14:52]$$

$$y = a + bx_4 + cx_4^2 + dx_4^3 + ex_4^4 \quad . \ . \ . \ . \ [14:53]$$

Then an interpolated value, x, from the following table in which all values except x are known, is the required answer.

$$\begin{array}{ccc} y_3 & \rightleftharpoons & x_3 \\ y & \rightleftharpoons & x \\ y_4 & \rightleftharpoons & x_4 \end{array}$$

Generally, if y is not intermediate between y_3 and y_4, further iterations are desirable before interpolating to obtain the final value of x.

If $f(x) = 0$ is to be solved for x, it will frequently be possible to write

$$f(x) = \phi(x) - \psi(x) = 0 \quad . \ . \ . \ . \ . \ [14:54]$$

in which $\phi(x)$ and $\psi(x)$ are not difficult to plot. The intersection, or intersections, of the two curves give values of x for which $f(x) = 0$. After a point of intersection has been approximately located from the graphs, select two values of x_1 and x_2 upon opposite sides of this point and as close to it as possible and compute $\phi(x_1)$, $\phi(x_2)$, $\psi(x_1)$, and $\psi(x_2)$. We desire a value x, between x_1 and x_2 such that $\phi(x) = \psi(x)$. If the points x_1 and x_2 are sufficiently close, the curves in this region may be represented by two intersecting straight lines, and from similar triangles it is readily derived that

$$x = \frac{(x_2 - x_1)\,[\phi(x_1) - \psi(x_1)]}{-\phi(x_2) + \phi(x_1) - \psi(x_1) + \psi(x_2)} + x_1 \quad [14:55]$$

A substitution of x into $\phi(x)$ and $\psi(x)$ will disclose the accuracy of the solution.

If $f(x)$ can be written as

$$f(x) = \phi(x)\psi(x) \quad . \ . \ . \ . \ [14:56]$$

we can employ logarithms, defining two new functions

$$P(x) = \log \phi(x) - \log f(x) \qquad [14:57]$$

$$Q(x) = -\log \psi(x)$$

which are related as in [14:54].

$$P(x) - Q(x) = 0$$

Solving as before we obtain x which satisfies [14:56].

THE NUMERICAL SOLUTION OF COMPLICATED SIMULTANEOUS EQUATIONS

The numerical solution of simultaneous equations which cannot readily be written in the form $y = f(x)$, can frequently be rapidly accomplished by iterative processes. If two equations are functions of x and y, they can generally be written in the following form in a number of ways:

$$x = f(x,y) \;.\;.\;.\;.\;.\;.\; [14:58]$$

$$y = F(x,y) \;.\;.\;.\;.\;.\;.\; [14:59]$$

An iterative process upon these equations may, or may not, converge in the neighborhood of some solution, x_0y_0. Graphs of [14:58] and [14:59] reveal, let us say, that $x = x_1$ and $y = y_1$ is an approximate solution. A second approximation is x_2,y_2, given by

$$x_2 = f(x_1, y_1) \;.\;.\;.\;.\;. \quad [14:60]$$

$$y_2 = F(x_2 \;\; y_1) \;.\;.\;.\;.\;. \quad [14:61]$$

A third approximation is x_3,y_3.

$$x_3 = f(x_2, y_2) \;.\;.\;.\;.\;. \quad [14:62]$$

$$y_3 = F \; x_3 \;\; y_2 \;.\;.\;.\;.\;. \quad [14:63]$$

This process is to be continued until two suc-

cessive sets of values differ negligibly. If convergence is not present, a different segregation of terms may be made so that we have $x = \phi(x,y)$ and $y = \psi(x,y)$ in lieu of [14:58] and [14:59]. Upon successive iteration these may lead to convergence. For iteration in the case of two such functions, say [14:58] and [14:59], to be convergent it is necessary that

$$\left|\frac{\partial f}{\partial x}\right| + \left|\frac{\partial F}{\partial x}\right| < 1 \quad . \; . \; . \; . \quad [14{:}64]$$

and

$$\left|\frac{\partial f}{\partial y}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \quad . \; . \; . \; . \quad [14{:}65]$$

in the neighborhood of $x_0 y_0$. This test for convergence can be made prior to iteration, or omitted if one prefers to discover the outcome by trial.

SECTION 8. TRANSFORMING RANK AND PERCENTAGE POSITIONS INTO QUANTITATIVE SCORES

Normalizing a distribution: In many situations in which a rank order, a percentage, or proportion position is given it is desired to treat the data quantitatively. In general, if p_1, p_2, and p_3 are three proportions, the numerical values of the differences $(p_1 - p_2)$ and $(p_2 - p_3)$ are an inaccurate quantitative statement of the underlying variables which yielded the three proportions. For example, if in a given trait, individual *A* exceeds p_1 of the members of his group, individual *B* exceeds p_2 and individual *C* exceeds p_3, and if the distribution of talent in the group is normal, the correct relationship is not $(p_1 - p_2)$ to $(p_2 - p_3)$, but $(x_1 - x_2)$ to $(x_2 - x_3)$ in which the *x*'s are the deviates in a normal distribution corresponding to the *p*'s. If, for each *p*, an *x* is abstracted from a table of a unit

normal distribution, the variable thus obtained will have many properties which fit it for quantitative study. This process is called normalizing a distribution. There obviously should be some plausibility in the assumption of near normality of the underlying trait before it is done. Even so, the x-variables resulting do not have all the characteristics of the quantitative measure in the case of a normally distributed trait.

If a test is equally reliable throughout a certain range, scores received by two individuals taking the test, whether low, intermediate, or high, have the same standard error which appropriately is called the standard error of measurement. In general, two proportions, p_1 and p_2, have unequal sampling errors, for, as given by [13:107], these standard errors are $\sqrt{p_1q_1/N}$ and $\sqrt{p_2q_2/N}$ respectively. Also the equivalent x_1 and x_2 deviates have unequal sampling errors, for as derived from [4:03], these standard errors are $\sqrt{p_1q_1/Nz_1^2}$ and $\sqrt{p_2q_2/Nz_2^2}$ respectively. Though the normalizing of ranks, or a series of proportions, or of percentages is known frequently to be very useful, nevertheless a transformation of such ranks, proportions, or percentages into scores which are equally reliable will clearly have value in cases where a single concept of the reliability of a score due to sampling is crucial in interpretation.

Transforming rank or percentage position into equally reliable deviation scores: Let us consider a transformation of p into x, such that σ_x (the standard error of each x, not the standard deviation of the distribution of the x's) is independent of the value of p,— that is, $\sigma_x = \sqrt{c/N}$, in which c is a constant and N the size of the sample yielding p. (An illustration follows

in which N is the number of items comprising a test.) The variable p is the observed value whose standard error is known, namely $\sqrt{pq/N}$, and the standard error which attaches to x is to be derived from that which attaches to p. Let

$$x = f(p) \quad \ldots\ldots\ldots\ldots\ldots \quad [14{:}66]$$

$$dx = f'(p)dp, \quad \ldots\ldots\ldots\ldots \quad [14{:}67]$$

and immediately

$$\sigma_x = f'(p)\sigma_p = f'(p)\sqrt{\frac{pq}{N}} = \sqrt{\frac{c}{N}}. \quad \ldots \quad [14{:}68]$$

or

$$f'(p) = \sqrt{\frac{c}{p-p^2}} \quad \ldots\ldots\ldots\ldots \quad [14{:}69]$$

Integrating

$$x = (2 \sin^{-1} \sqrt{p} - \frac{\pi}{2}) \sqrt{c} + k \quad \ldots\ldots \quad [14{:}70]$$

If k, the constant of integration, be set equal to 0, and c set equal to 1, we have

$$x = 2 \sin^{-1}\sqrt{p} - \frac{\pi}{2} \quad \ldots\ldots\ldots\ldots \quad [14{:}71]$$

Then x takes values from $-\pi/2$ to $\pi/2$ as p takes values from 0 to 1, and σ_x, no matter the value of x, is a function of the size of the sample only and is equal to $1/\sqrt{N}$. This σ_x is the standard error of each x and not the standard deviation of the distribution of x's, which takes a range of values depending upon the distribution of the p's.

In certain problems it is desirable to make an analysis of the variance of a set of independent proportions, but the proportions, based

upon the same numbers of cases, are different and accordingly unequally reliable. If each proportion is transformed into an x by [14:71], the resulting measures are independent and equally reliable so that their further study by analysis of variance methods is simple and accurate.

As an illustration of a problem which might well use this transformation, consider a fundamental operations in arithmetic test, consisting of N problems, of substantially equal difficulty, administered without time limit and scored for speed of completion and again for accuracy of computation. The accuracy score is p, the proportion of answers that are correct and as explained it has unequal reliability from subject to subject. If for each p we employ [14:71] or Table XIV B to obtain x, we now have a score with a constant standard error $1/\sqrt{N}$. This happy outcome has been accomplished at a certain expense. The distribution of the x's is presumably not a correct representation of the true distribution of underlying arithmetical computation ability. Specifically, had this true distribution been normal neither the distribution of p's nor that of the x's would be normal. How serious this is becomes an appropriate topic for investigation in each particular instance.

The [14:71] transformation may be useful when the original data consists of ranks, or equivalent percentages, or proportions. If N subjects are ranked in an order of merit, the proportion of cases falling below a person is his p-score. In computing this the proportion represented by the person himself is split into halves, one-half being added to the proportion below him and one-half to the proportion above him; thus the person ranking lowest has a p-score $= .5/N$; the second lowest a $p = 1.5/N$; the third lowest a $p = 2.5/N$; etc. These p-scores form a rectangular distri-

TABLE XIV B

Of $x = 2 \sin^{-1} \sqrt{p} - \frac{\pi}{2}$

(negative for $p < .5$ and positive for $p > .5$)			(negative for $p < .5$ and positive for $p > .5$)			(negative for $p < .5$ and positive for $p > .5$)		
p	x	p	p	x	p	p	x	p
.00	1.5708	1.00	.17	.7208	.83	.34	.3257	.66
.01	1.3705	.99	.18	.6945	.82	.35	.3047	.65
.02	1.2870	.98	.19	.6687	.81	.36	.2838	.64
.03	1.2226	.97	.20	.6435	.80	.37	.2631	.63
.04	1.1681	.96	.21	.6187	.79	.38	.2424	.62
.05	1.1198	.95	.22	.5944	.78	.39	.2218	.61
.06	1.0759	.94	.23	.5704	.77	.40	.2014	.60
.07	1.0353	.93	.24	.5469	.76	.41	.1810	.59
.08	.9973	.92	.25	.5236	.75	.42	.1607	.58
.09	.9614	.91	.26	.5007	.74	.43	.1405	.57
.10	.9273	.90	.27	.4780	.73	.44	.1203	.56
.11	.8947	.89	.28	.4556	.72	.45	.1002	.55
.12	.8633	.88	.29	.4334	.71	.46	.0801	.54
.13	.8331	.87	.30	.4115	.70	.47	.0600	.53
.14	.8038	.86	.31	.3898	.69	.48	.0400	.52
.15	.7754	.85	.32	.3683	.68	.49	.0200	.51
.16	.7478	.84	.33	.3469	.67	.50	.0000	.50

bution which we may treat as continuous if N is at all substantial. Whereas in the arithmetic fundamentals test the p's may yield a unimodal distribution covering a small range, in the present case the p's extend uniformly from $p = 0$ to $p = 1$. We now know the form of the distribution of p and if the [14:71] transformation is made we can determine the form of distribution of x. The equation of the distribution of p, so written that the total area is 1, is

$$z_p = 1 \quad (\text{from } p=0 \text{ to } p=1) \qquad [14:72]$$

Differentiating [14:71]

$$dx = \frac{1}{\sqrt{p-p^2}}\,dp \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [14:73]$$

Let the number of cases in the interval dx be f_x and the number in the interval dp be f_p.

$$z_x = \frac{f_x}{dx} \quad \text{The ordinate in the } x\text{-distribution} \quad [14:74]$$

$$z_p = \frac{f_p}{dp} = 1 \quad \text{The ordinate in the } p\text{-distribution} \quad [14:75]$$

The element of area $z_x dx$ equals the corresponding element of area $z_p dp$, that is $z_x dx = dp$. Utilizing [14:73] we have

$$z_x = \sqrt{p-p^2} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [14:76]$$

Obtaining p from [14:71] and substituting we get

$$z_x = \sin\left(\frac{x}{2} + \frac{\pi}{4}\right) \cos\left(\frac{x}{2} + \frac{\pi}{4}\right) \ldots$$

$$\text{The sine-cosine distribution} \quad \frac{-\pi}{2} \leqq x \leqq \frac{\pi}{2} \quad [14:77]$$

This is a symmetrical distribution, with limits in x of $-\pi/2$ and $\pi/2$, of total area 1, having a mean of zero, a variance of $(\pi^2-8)/4$, a Pearson β_1 of 0, a Pearson β_2 of 2.19375

$$\left(= \frac{\pi^4 - 48\pi^2 + 384}{\pi^4 - 16\pi^2 + 64}\right)$$

an area between any two values as given by [14:78], and a standard error of every x, if derived from a p with standard error of $\sqrt{pq/N}$, (which, presumably will seldom be the case), of $1/\sqrt{N}$. The area between x_1 and x_2 is given by

$$\int_{x_1}^{x_2} z_x\, dx = \left[\sin^2\left(\frac{x}{2} + \frac{\pi}{4}\right)\right]_{x_1}^{x_2} = \sin^2\left(\frac{x_2}{2} + \frac{\pi}{4}\right) - \sin^2\left(\frac{x_1}{2} + \frac{\pi}{4}\right) \qquad \frac{-\pi}{2} \leq x \leq \frac{\pi}{2} \qquad [14:78]$$

If a table of sines for circular arguments is not available, x, may be expressed in terms of degrees employing the relationship

$$x \text{ radians} = \frac{180}{\pi} \text{ degrees, or}$$

$$57.29577\ 951x = \text{corresponding number of degrees} \qquad [14:79]$$

$$.01745\ 32925 \text{ (degrees)} = \text{corresponding number of radians} \qquad [14:80]$$

SECTION 9. THE SQUARE ROOT TRANSFORMATION (See Bartlett 1936)

In the situation wherein independent variables are normally distributed a sum (or weighted sum by introducing weights), as given by [14:81], (see Chapter VIII, Section 10).

$$X = X_1 + X_2 + \cdot \cdot \cdot + X_k \ . \ . \ . \ . \qquad [14:81]$$

is distributed normally, and such an estimate as X/k is of the mean of the right hand member X's,

is an efficient estimate, if these X's are equally reliable. The analysis of variance proceeds via the equation,

$$V_x = V_1 + V_2 + \ldots + V_k.$$

If the independent X's are Poisson distributed a more efficient prediction equation than one predicting X/k is one predicting $\sqrt{X/k}$, thus, if $Y_1 = \sqrt{X_1}$, $Y_2 = \sqrt{X_2}$, . . . , $Y_k = \sqrt{X_k}$,

$$Y = Y_1 + Y_2 + \ldots + Y_k$$

$$V(Y) = V(Y_1) + V(Y_2) + \cdots + V(Y_k) \quad [14:82]$$

Though this is not a totally adequate transformation* (See Cochran 1940) it has such merit as to suggest its use when the original variables are Poisson distributed and an analysis of variance is desired.

SECTION 10. CUBE ROOT TRANSFORMATION

The Wilson-Hilferty transformation is a linear function of a cube root. It very nearly normalizes the χ^2 distribution. As employed in Chapter IX, Section 5 it is the basis for very nearly normalizing variance ratio distributions. As these distributions are to be found in all realms of social, biological, and physical statistics the cube root transformation may be expected to have wide utility.

SECTION 11. CERTAIN PROPERTIES OF DIFFERENCES IN TABLED ENTRIES

We employ the notation of Chapter XIII, Section 12.

If in a table all t's except t_0 are correct and t_0 is too large by the amount ϵ, the errors in the differences take the pattern:

*Cochran refines his $\sqrt{X}$ values, by an adjustment which is small except in the case of an X that equals 0.

						$\Delta^{VI}_{-6}+\epsilon$
					$\Delta^{V}_{-5}+\epsilon$	
				$\Delta^{IV}_{-4}+\epsilon$		$\Delta^{VI}_{-5}-6\epsilon$
			$\Delta^{III}_{-3}+\epsilon$		$\Delta^{V}_{-4}-5\epsilon$	
		$\Delta^{II}_{-2}+\epsilon$		$\Delta^{IV}_{-3}-4\epsilon$		$\Delta^{VI}_{-4}+15\epsilon$
	$\Delta^{I}_{-1}+\epsilon$		$\Delta^{III}_{-2}-3\epsilon$		$\Delta^{V}_{-3}+10\epsilon$	
$t_0+\epsilon$		$\Delta^{II}_{-1}-2\epsilon$		$\Delta^{IV}_{-2}+6\epsilon$		$\Delta^{VI}_{-3}-20\epsilon$
	$\Delta^{I}_{0}-\epsilon$		$\Delta^{III}_{-1}+3\epsilon$		$\Delta^{V}_{-2}-10\epsilon$	
		$\Delta^{II}_{0}+\epsilon$		$\Delta^{IV}_{-1}-4\epsilon$		$\Delta^{VI}_{-2}+15\epsilon$
			$\Delta^{III}_{0}-\epsilon$		$\Delta^{V}_{-1}+5\epsilon$	
				$\Delta^{IV}_{0}+\epsilon$		$\Delta^{VI}_{-1}-6\epsilon$
					$\Delta^{V}_{0}-\epsilon$	
						$\Delta^{VI}_{0}+\epsilon$

It will be noticed that the largest error in the even Δ's is opposite the tabled entry which is in error. When an error in some t value is suspected it is frequently possible to locate it by thus studying the differences.

The following formula, noted by Cosens (1944) is particularly useful to check an entry, t_0, which one questions, by determining what this entry would be based on the six neighboring entries.

$$t_0 = .75(t_{-1}+t_1) -.30(t_{-2}+t_2) + .05(t_{-3}+t_3) -.05\,\Delta^{VI}t_{-3} \qquad [14:83]$$

When the recorded t_0 is presumably in error the value $\Delta^{VI}t_{-3}$ will, presumably, be greatly in error. If neighboring portions of the table warrant the belief that Δ^{VI} is negligibly small, the $\Delta^{VI}t_{-3}$ term in [14:83] may be dropped.

SECTION 12. EXPANDING A TABLE BY INTERPOLATING VALUES

Let us be given a table with entries t for equally spaced arguments a and let it be desired to interpolate $(n-1)$ additional values between a t value and a neighboring t value. For example, if the arguments run .00, .01, .02, . . . with corresponding tabled entries, it might be desired to expand the table so that the arguments run .000, .001, .002, . . . with corresponding tabled entries. The interval in the expanded table is 1/10 that in the original table. In this case $n = 10$ and 9 additional values are to be interpolated between existing values. The tabled entries have, of necessity, an error which may be as large as 1/2 in the last decimal place recorded, so these tabled entries should be accurate to one or more, preferably more, decimal places farther than is required in the interpolated values.

Differences in the original table are computed to the first order that becomes negligibly small. We will assume this to be the sixth order, so that $\Delta^{VI} = 0$, and illustrate the solution for this case. If Δ's represent the differences in the original table and δ's in the expanded table, the relationships between them are given herewith. For simplicity the subscript 0, or any other subscript which remains constant, which attaches to every δ and Δ, has been omitted.

$$\delta^{VI} = \frac{\Delta^{VI}}{n^6} \doteq 0 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad [14:84]$$

$$\delta^{V} = \frac{\Delta^{V}}{n^5} \quad \text{[14:85]}$$

$$\delta^{IV} = \frac{\Delta^{IV}}{n^4} - 2(n-1)\frac{\Delta^{V}}{n^5} \quad \text{[14:86]}$$

$$\delta^{III} = \frac{\Delta^{III}}{n^3} - \frac{3(n-1)}{2}\frac{\Delta^{IV}}{n^4} + \frac{(n-1)(7n-5)}{4}\frac{\Delta^{V}}{n^5} \quad \text{[14:87]}$$

$$\delta^{II} = \frac{\Delta^{II}}{n^2} - (n-1)\frac{\Delta^{III}}{n^3} + \frac{(n-1)(11n-7)}{12}\frac{\Delta^{IV}}{n^4}$$

$$- \frac{(n-1)(2n-1)(5n-3)}{12}\frac{\Delta^{V}}{n^5} \quad \text{[14:88]}$$

$$\delta^{I} = \frac{\Delta^{I}}{n^1} - \frac{(n-1)}{2}\frac{\Delta^{II}}{n^2} + \frac{(n-1)(2n-1)}{3!}\frac{\Delta^{III}}{n^3}$$

$$- \frac{(n-1)(2n-1)(3n-1)}{4!}\frac{\Delta^{IV}}{n^4}$$

$$+ \frac{(n-1)(2n-1)(3n-1)(4n-1)}{5!}\frac{\Delta^{V}}{n^5} \quad \text{[14:89]}$$

Having the δ's, the interpolated values are readily obtained. A ready check for accuracy is available by employing overlapping regions and

computing certain of the interpolated values twice. The preceding formulas serve by dropping off the higher order terms if some Δ of lower order than Δ^{VI} approximates zero. If it is a higher order than Δ^{VI} that approximates zero, similar formulas to the preceding can be derived.

SECTION 13. TRIGONOMETRIC FUNCTIONS OF SUMS AND DIFFERENCES

Given sin, cos, and tan of θ,

$$\sin\frac{\theta}{2} = \sqrt{\frac{1}{2}(1-\cos\theta)};\quad \cos\frac{\theta}{2} = \sqrt{\frac{1}{2}(1+\cos\theta)};$$

$$\tan\frac{\theta}{2} = \frac{\sin\theta}{1+\cos\theta} \qquad [14:90]$$

$$\sin(2\theta) = 2\sin\theta\cos\theta;\quad \cos(2\theta) = \cos^2\theta - \sin^2\theta \qquad [14:91]$$

$$\tan(2\theta) = \frac{2\tan\theta}{1-\tan^2\theta} \;.\;.\;. \qquad [14:92]$$

Given sin, cos, and tan of θ and of ϕ,

$$\sin(\theta\pm\phi) = \sin\theta\cos\phi \pm \cos\theta\sin\phi \;.\;.\;.\;. \qquad [14:93]$$

$$\cos(\theta\pm\phi) = \cos\theta\cos\phi \mp \sin\theta\sin\phi \;.\;.\;.\;. \qquad [14:94]$$

$$\tan(\theta\pm\phi) = \frac{\tan\theta \pm \tan\phi}{1 \mp \tan\theta\tan\phi} \;.\;.\;.\;.\;.\;.\;.\;. \qquad [14:95]$$

SECTION 14. SPACE OF TWO DIMENSIONS

A few *properties of a straight line*. There are various forms for the equation.

$ax + by + c = 0$ The general form [14:96]

$y = a + bx$ In which a is the intercept on the y axis and b the slope of the line. [14:97]

$$\frac{y}{a} + \frac{x}{b} = 1 \qquad \text{In which } a \text{ and } b \text{ are the intercepts on the } y \text{ and } x \text{ axes respectively} \qquad [14:98]$$

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1} \qquad \text{In which } (x_1, y_1) \text{ and } (x_2, y_2) \text{ are two points through which the line passes.} \qquad [14:99]$$

The perpendicular distance of the point (x_1, y_1) from the line [14:96] is

$$d = \frac{ax_1 + by_1 + c}{\sqrt{a^2 + b^2}} \qquad \text{In which the radical takes the sign opposite to that of } c. \qquad [14:100]$$

Rotation of axes in two dimensional space: Let x_1 and x_2 be the original variables and y_1 and y_2 the variables after rotation. If the axes rotate counter clock-wise through the angle θ, the transforming equations are

$$y_1 = x_1 \cos\theta + x_2 \sin\theta \quad \ldots\ldots [14:101]$$

$$y_2 = -x_1 \sin\theta + x_2 \cos\theta \quad \ldots\ldots [14:102]$$

The reverse equations, expressing the same relationship, are

$$x_1 = y_1 \cos\theta - y_2 \sin\theta \quad \ldots\ldots [14:101a]$$

$$x_2 = y_1 \sin\theta + y_2 \cos\theta \quad \ldots\ldots [14:102a]$$

Correlation functions of rotated variables: Let x_1, x_2, and x_3 be three correlated variables with means of zero, variances of V_1, V_2, and V_3, and covariances c_{12}, c_{13}, and c_{23}. Let x_3 remain unchanged, but x_1 and x_2 transformed into y_1 and y_2 by the rotation given by [14:101] and [14:102]. Then

$$V(y_1) = V_1\cos^2\theta + V_2\sin^2\theta + 2\,c_{12}\sin\theta\,\cos\theta \quad [14{:}103]$$

$$V(y_2) = V_1\sin^2\theta + V_2\cos^2\theta - \,c_{12}\sin\theta\,\cos\theta \quad [14{:}104]$$

$$V(y_1) + V(y_2) = V_1 + V_2 \quad \ldots\ldots\ldots \quad [14{:}105]$$

$$c(y_1y_2) = c_{12}(\cos^2\theta - \sin^2\theta) - (V_1 - V_2)\sin\theta\,\cos\theta \quad \ldots\ldots \quad [14{:}106]$$

$$c(y_1x_3) = c_{13}\cos\theta + c_{23}\sin\theta \quad \ldots\ldots \quad [14{:}107]$$

$$c(y_2x_3) = -c_{13}\sin\theta + c_{23}\cos\theta \quad \ldots\ldots \quad [14{:}108]$$

$$[c(y_1x_3)]^2 + [c(y_2x_3)]^2 = c_{13}^2 + c_{23}^2 \quad [14{:}109]$$

If the rotation in the x_1, x_2 plane is to major and minor axes, or such as to make $V(y_1)$ a maximum, it of necessity at the same time makes $V(y_2)$ a minimum and the covariance, $c(y_1y_2)$, zero. In this case the angle θ is given by

$$\tan(2\theta) = \frac{2c_{12}}{V_1 - V_2} \quad \ldots\ldots\ldots \quad [14{:}110]$$

Tables to facilitate such rotations, giving $\sin\theta$, $\cos\theta$, $\sin^2\theta$, $\cos^2\theta$, $\sin\theta\,\cos\theta$, and $2\sin\theta\,\cos\theta$ for argument θ or argument $\tan(2\theta)$ are given in Kelley (1935).

SECTION 15. SPACE OF THREE OR MORE DIMENSIONS

The general form of the equation of a plane in three-dimensional space is

$$ax + by + cz + d = 0 \quad \ldots\ldots\ldots \quad [14{:}111]$$

The intercept form in which a, b, and c are the intercepts on the x, y, and x axes, is

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1 \quad \ldots\ldots\ldots\ldots [14:112]$$

The distance of the plane [14:111] from the point (x_1, y_1, z_1) is

$$d = \frac{ax_1 + by_1 + cz_1 + d}{\sqrt{a^2 + b^2 + c^2}} \quad \ldots\ldots [14:113]$$

in which the sign of the radical is opposite to that of d.

Two non-parallel planes intersect in a straight line. Accordingly the equations of two planes define a line.

Three planes, in general, intersect in a point and the equations of three planes define a point.

For the angle between two lines, we let the lines be defined by

$$\frac{x-x_1}{a_1} = \frac{y-y_1}{b_1} = \frac{z-z_1}{c_1} \quad \ldots\ldots [14:114]$$

and

$$\frac{x-x_2}{a_2} = \frac{y-y_2}{b_2} = \frac{z-z_2}{c_2} \quad \ldots\ldots [14:115]$$

then

$$\cos\theta = \frac{a_1a_2 + b_1b_2 + c_1c_2}{\sqrt{a_1^2 + b_1^2 + c_1^2}\sqrt{a_2^2 + b_2^2 + c_2^2}} \quad \ldots [14:116]$$

in which θ is the angle between the two lines.

The sine of the angle between the straight line [14:114] and the plane [14:111] is

$$\sin\theta = \frac{a_1a + b_1b + c_1c}{\sqrt{a_1^2 + b_1^2 + c_1^2}\sqrt{a^2 + b^2 + c^2}} \quad \ldots [14:117]$$

SECTION 16. LAGRANGE MULTIPLIERS

These multipliers provide a method for finding the maximum or minimum of a function when one or more conditions are imposed upon the variables in the function. The general statement is as follows:

Let f be the function to be maximized, or minimized. Let $\phi_1=0;\ .\ .\ .\ \phi_k=0;$ be the conditions imposed upon the variables. $f,\ \phi_1,\ \phi_2,\ .\ .\ .\ ,\ \phi_k$ are functions of $x_1,\ x_2,\ .\ .\ .\ ,\ x_n,$ Write

$$\psi = f + \lambda_1\phi_1 + \lambda_2\phi_2 + \ . \ . \ . \ + \lambda_k\phi_k \qquad [14:118]$$

Obtain the n partial derivates and set equal to zero.

$$\frac{\partial \psi}{dx_1} = 0$$

$$\frac{\partial \psi}{dx_2} = 0 \qquad . \ . \ . \ . \ . \qquad [14:119]$$

$$\vdots$$

$$\frac{\partial \psi}{dx_n} = 0$$

These n equations, together with the k equations, $\phi_1=0,\ \phi_2=0,\ .\ .\ .\ .\ ,\ \phi_k=0$ provide a set of $(n+k)$ equations from which the λ's may be eliminated and the maximizing, or minimizing, values of the x's obtained.

To illustrate: Let $f = \dfrac{1}{2}xy$ = the area of a triangle with base x and altitude y, and let it be further imposed that $x + y^2 = 9$. Required to find the dimensions yielding a triangle with maximum area. We have

$$f = \frac{1}{2} xy \quad \ldots\ldots\ldots\ldots \quad [a]$$

$$\phi_1 = x + y^2 - 9 = 0 \quad \ldots\ldots\ldots\ldots \quad [b]$$

$$\psi = \frac{1}{2} xy + \lambda_1(x+y^2-9) \quad \ldots\ldots\ldots\ldots \quad [c]$$

$$\frac{\partial\psi}{dx} = \frac{1}{2} y + \lambda_1 = 0 \quad \ldots\ldots\ldots\ldots \quad [d]$$

$$\frac{\partial\psi}{dy} = \frac{1}{2} x + 2\lambda_1 y = 0 \quad \ldots\ldots\ldots\ldots \quad [e]$$

Solving [b], [d], and [e] simultaneously we obtain $x = 6$ and $y = \sqrt{3}$ for the required dimensions of the triangle.

SECTION 17. GROWTH CURVES

a, b, and c are constants and x is time, or age. The law of organic growth

$$y = a\, e^{bx} \quad \ldots\ldots\ldots\ldots \quad [14:120]$$

Makeham's force of mortality curve

$$y = c + a\, e^{bx} \quad \ldots\ldots\ldots\ldots \quad [14:121]$$

The simple logistic growth curve

$$y = \frac{a}{1 + be^{-cx}} \quad \ldots\ldots\ldots\ldots \quad [14:122]$$

The Gomperz growth curve

$$y = a^{b^x} \quad \ldots\ldots\ldots\ldots \quad [14:123]$$

A growth senescence curve

$$y = \frac{a}{e^{-bx} + e^{cx}} \quad \ldots\ldots\ldots\ldots \quad [14:124]$$

SECTION 18. THE BINOMIAL THEOREM

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2$$

$$+ \frac{n(n-1)(n-2)}{3!} a^{n-3}b^3 + \dots$$

$$+ \frac{n(n-1)(n-2) \dots (n-n+1)}{n\,!} a^{n-n}b^n \quad [14:125]$$

The expansion of the polynomial $(a+b+c+ \dots +k)^n$ consists of terms the sum of whose exponents is always equal to n. If for some term the exponent of a is α, of b, β, of c, γ, . . . of k, κ, where of course $\alpha + \beta + \gamma + \dots + \kappa = n$, this term is given by

$$\frac{n!}{\alpha!\ \beta!\ \gamma!\ \dots\ \kappa!} a^{\alpha}b^{\beta}c^{\gamma} \dots k^{\kappa} \dots [14:126]$$

SECTION 19. STIRLING'S APPROXIMATION TO THE FACTORIAL

$$\ln(N!) = (N+\frac{1}{2}) \ln(N) + \frac{1}{2} \ln(2\pi) - N$$

$$+ \frac{B_1}{2!\ N} - \frac{2!\ B_2}{4!\ N^3} + \frac{4!\ B_3}{6!\ N^5} \quad [14:127]$$

The B's are Bernoullian numbers.

SECTION 20. THE SUM OF THE POSITIVE POWERS OF THE FIRST k NUMBERS

$$1^p + 2^p + \dots + k^p = \frac{k^{p+1}}{p+1} + \frac{k^p}{2} + \frac{pk^{p-1}B_1}{2!}$$

[14:128]

$$\frac{p(p-1)(p-2)k^{p-3}B_2}{4!} + \frac{p(p-1)(p-2)(p-3)(p-4)k^{p-5}B_3}{6!} - \dots$$

the series terminating with the term in which the exponent of k is 1 or 2. The B's are Bernoullian numbers.

SECTION 21. FOURIER SERIES

$$\bar{y} = a_0 + a_1 \sin x + b_1 \cos x + a_2 \sin 2x + b_2 \cos 2x + \ldots \qquad [14:129]$$

This is important in fitting curves to data showing periodicity. This is treated in E. T. Whittaker and G. Robinson, *THE CALCULUS OF OBSERVATIONS*, (1924) Chapter X.

SECTION 22. TAYLOR'S THEOREM

$f(x)$ may be expanded in the neighborhood of $x = a$ thus

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!} f''(a) + \ldots$$

$$\frac{(x-a)^{n-1}}{(n-1)\,!} f^{n-1}(a) + R \qquad [14:130]$$

in which R is the remainder after n terms, and the successive f-primes are the successive derivates of the function evaluated for $x = 0$. If the limit of R is zero as $n \to \infty$ we have a Taylor's series, and this is a convergent series. If $a=0$ in a Taylor's series of one variable we have a Maclaurin series. Taylor's theorem also applies, with certain necessary modifications, to functions of more than one variable.

SECTION 23. THE EULER-MACLAURIN FORMULA FOR EVALUATING A DEFINITE INTEGRAL

Let k be the number of intervals of uniform base in the region from a to $a + ki$. Then the

integral of $f(x)\,dx$ is given by

$$\frac{1}{i}\int_{a}^{a+ki} f(x)dx = \frac{1}{2} f(a) + f(a+i) + f(a+2i) + \ldots + f(a+\overline{k-1}\ i) + \frac{1}{2} f(a+ki) - \frac{iB_1}{2!}\left[f'(a+ki)-f'(a)\right] + \frac{i^3B_2}{4!}\left[f'''(a+ki)-f'''(a)\right] - \frac{i^5B_3}{6!}\left[f^5(a+ki)-f^5(a)\right] + \text{etc.} \qquad [14:131]$$

in which the B's are Bernoullian numbers.

CHAPTER XV

STATISTICAL TABLES

SECTION 1. SELECTED REFERENCES

The tables given in later sections are likely to be too abbreviated for refined statistical work. They should prove adequate for training purposes and for preliminary work. Much more extensive tables of the sorts here given are available in *The Kelley Statistical Tables*, in press, 1947.

The accompanying very limited and select references indicate the sources of what are deemed to comprise certain tables important to the statistician. The arrangement is chronological, except where several tables appear under a single institutional authorship.

1814. BARLOW'S TABLES. 1814 and later. Edition of 1930 here referred to. [Generally 8 significant figures in tabled values, with an argument of 4 figures] "Squares, cubes, square roots, cube roots, and reciprocals of all integer numbers up to 10,000." The square root of N and of $10N$ are given. The cube root of N is given, but not that of $10N$ nor that of $100N$. A table of powers up to the tenth of all integers from 1

to 100, and of powers up to the 20th of all integers from 1 to 10, is given.

1910. B. O. Peirce. A SHORT TABLE OF INTEGRALS. Second edition, 1910,- 144 pages. This is a handy and concise source for 938 formulas: integrals, auxiliary trigonometric, hyperbolic, elliptic functions, Bessel's functions, series, products, derivatives, etc. Also contains brief tables of the probability integral, elliptic integrals, hyperbolic functions, natural and common logarithms, and trigonometric functions. Third edition revised by **W. F. Osgood**, 1929,-156 pages.

1914. Karl Pearson, editor. **TABLES FOR STATISTICIANS AND BIOMETRICIANS, Part I.** (See also Part II, 1931; Cambridge University Press, Tracts for Computers, 1919-37; and, Biometrika Publications, 1934-38).

Table II (W. F. Sheppard) [7 place consequent] *Area and ordinate of the normal curve in terms of the abscissa.* Interval is .01 σ. (1/2 (1+α) = p and z = z of Table XV C herein.)

Table XII (W. Palin Elderton) [6 places] *Tables for testing goodness of fit.* These are χ^2 tables. The interval in χ^2 is very coarse, being 1. It is important to note that n' as used in these tables is one greater than d. o. f. The n' arguments are 3(1) 30.

Table XXVI (W. Palin Elderton) Tables to assist the calculation or the ordinates of [*a type III*] *curve.*

Table XXVII (W. Palin Elderton) Tables of powers of natural numbers 1 to 100. Also *Table XXVIII Sums of powers of natural numbers 1 to 100.*

Table XXX (P. F. Everitt) [4 place] *Tables to facilitate the determination of tetrachoric correlation.* r =.80, 85, 90, 95, and 1.00. For more complete tabulations see *Pearson's Tables,* Pt. II, Tables VIII and IX.

Table XXXI (J. H. Duffell) [7 places] *Logarithms of the gamma function . . . from p = 1 to p =* 2 [interval .01].

XXXV A *Diagram to determine the type of a frequency distribution from a knowledge of the constants* β_1 *and* β_2.

Tables XXXVII to XLVII (A. J. Rhind) Probable errors of frequency constants connected with the Pearson system of curves.

Table XLIX (Julia Bell) [Beginning with 7 place and ending with 11 place] *Logarithm of factorial n from n =* 1 *to* 1000.

Tables LI and LII concern Poisson distributions.

Table LIV (Alice Lee) Tables of the $G(r,\nu)$ *integrals.* $G(r,\nu) = \sin^r\theta\; e^{\nu\theta}\, d\theta$: Auxiliary functions leading to this integral are tabled for arguments $r = 1\ (1)\ 50$ and $\phi^\circ = 1\ (1)\ 45$,—ϕ being defined by $\tan\phi = \nu/r$.

1919-1939. Cambridge University Press Tracts for Computers, Karl Pearson, Editor. Department of Applied Statistics, University of London.

No. 1, 1919. (Eleanor Pairman) Tables of the Digamma and Trigamma Functions. The digamma function is defined $F(z) = d/dz\ \ln\Gamma(1+z)$, and the trigamma function is defined

$$F(z)\ \frac{d^2}{dz^2} \ln\Gamma(1+z)$$

F and F are tabled for argument $x = .00\ (.02)\ 20.00$.

No. 4, 1921. (Originally computed by *A. M. Legendre) Tables of the Logarithms of the Complete Gamma Function to Twelve Figures.* Log Γ a is tabled. Argument $a = 1.000\ (.001)\ 2.000$.

No. 8, 1922. (Egon S. Pearson) Table of the Logarithms of the Complete Gamma Function (for arguments 2 to 1200, i.e., beyond Legendre's range). Log $\Gamma(p)$ is tabled to 10 decimal

places. Argument p=2.0 (.1) 5.0 (.2) 70. (1) 1200.

No. 9, 1923. (John Brownlee) Log$\Gamma(x)$ from x=1 to 50.9 by intervals of .01. [7 decimal places].

No. 13, 1926. (James Henderson) Bibliotheca Tabularum Mathematicarum, being a descriptive catalog of mathematical tables. Part I, Logarithmic Tables.

Alexander John Thompson. Logarithmetica Britannica, being a standard table of logarithms to twenty decimal places. No. 11, numbers 90,000 to 100,000 [1924]; no. 14, numbers 80,000 to 90,000 [1927]; no. 16, numbers 40,000 to 50,000 [1928]; no. 17, numbers 50,000 to 60,000 [1931]; no. 18, numbers 60,000 to 70,000 [1923]; no. 19, numbers 10,000 to 20,000 [1934]; no. 20, numbers 70,000 to 80,000 [1935]; no. 21, numbers 30,000 to 40,000 [1937]. The part for numbers 20,000 to 30,000 announced in 1945 as "at press."

No. 15, 1927. (L. H. C. Tippett) Random Sampling Numbers. 10,400 four-figure numbers.

No. 23, 1938. (L. J. Comrie) Tables of $\tan^{-1} x$ *and* $\log(1+x^2)$.

No. 24, 1939. (M. G. Kendall and B. Babington Smith) Random Sampling Numbers,—Second Series. 100,000 digits grouped in twos and fours and in 100 separate thousands.

1923. James W. Glover. TABLES OF APPLIED MATHEMATICS IN FINANCE, INSURANCE, STATISTICS. Part I "Values of compound interest functions to 8 places of decimals and 7 place logarithms of values of compound interest functions." Part II "Values of life insurance and disability insurance functions." Part III "Values of probability and statistical functions." Part IV "Common logarithms of numbers from 1 to 100,000 to 7 places of decimals."

1930. L. R. Salvosa. *Tables of Pearson type III functions* and *Derivatives of Pearson type III curve.* Annals of Math. Stat., Vol. 1, 1930 [six decimal places].

Table I. Areas. Argument, t, is a standard

score deviate. $t = -4.99$ (.01) 9.99. Skewness argument, $\alpha_3 = \sqrt{\beta_1}$, is .0 (.1) 1.1.

Table II. Ordinates. Argument $t = -5.49$ (.01) 9.99. Skewness $\alpha_3 = 0$ (.1) 1.1.

Table III. First six derivatives. Argument $t = -9.9$ (.1) 14.9. Skewness $\alpha_3 = .0$ (.1) 1.1.

1931. Karl Pearson, Ed. TABLES FOR STATISTICIANS AND BIOMETRICIANS, Part II.

Table II (T. Kondo and E. M. Elderton) Abscissae, ordinates and ratios [z/p, z/q, p/z, q/z] to ten decimal places of the normal curve to each permille of frequency.

Table III (John P. Mills and B. H. Camp) [5 places] Ratio, area of tail to bounding ordinate, or [q/x] *for each percent. of deviate.*

Tables V-VII (Alice Lee) concern tetrachoric functions.

Tables VIII-IX (Alice Lee, Margaret Moul, Ethel M. Elderton, A. E. R. Church, E. C. Fieller, J. Pretorius, and K. Pearson) [6 place] "*For determining the volume of any quadrant or of any cell of a bivariate normal frequency distribution.*" The interval in both arguments is $.1\sigma$. A table for each of the following values of r is given: $r = -.95$ (.05) 1.00.

Tables XXI-XXII (L. H. C. Tippett) [7 and 6 place] "*Distribution of extreme individuals and of the range in samples from a normal population.*" Table XXI gives distributions for N = 3, 5, 10, 20, 30, 50, 100, (100) 1000. Table XXII gives mean range for N = 1 (1) 1000.

Table XXV (K. Pearson and B. Stoessiger) [7 place] *Probability integral for symmetrical curves; $\beta_1 = 0$; β_2 = 1 to 3 and 3 onwards.*

1932. Jack W. Dunlap and Albert K. Kurtz. HANDBOOK OF STATISTICAL NOMOGRAPHS, TABLES, AND FORMULAS.

Part I. Gives nomographs for obtaining probable errors, or standard errors, or standard deviations of means, sums, differences, frequency in a class, proportion in a class, upper or lower

quartile, quartile deviation, 10-90 percentile range, r, rank correlation coefficient, and also nomographs for biserial r, regression weights based upon biserial r, correction for attenuation, effect of range upon standard deviations and reliability coefficients, intelligence quotients, percentages, etc.

Part II:

Table 72. Squares, square roots [4 decimal places], *reciprocals* [5 and 6 significant figures], *and reciprocals of square roots* [6 significant figures]. Argument 1 (1) 1000.

Table 84. "$\sqrt{1-r^2}$" [4 places] Argument .000 (.001) .999.

Table 86. "$1-r^2$" [4 places] Argument .000 (.001) .999.

Table 88. Reliability of two to ten tests,— Spearman-Brown formula. [4 places] Argument r_1 = .00 (.01) .99.

Table 95. Functions used in rank correlation. d^2 tabled for differences in rank, by .5, from .5 to 75. $N(N^2-1)$ tabled for N = 1 (1) 100. Table also useful in getting N^3.

Table 98. Correlation coefficient "From the percentage of unlike signed pairs."

Part III is an extensive table of formulas as used by different authors.

1933. Leone Chesire, Milton Saffir, and L. L. Thurstone. COMPUTING DIAGRAMS FOR THE TETRACHORIC CORRELATION COEFFICIENTS. Consisting of 46 trivariate diagrams, which, when entered by means of the proportionate frequencies in two marginal totals and one cell frequency, yield tetrachoric r. Three decimal place accuracy may be expected.

1933-1935. Harold T. Davis. TABLES OF THE HIGHER MATHEMATICAL FUNCTIONS. Vol. I, 1933. Vol. II, 1935. There is given throughout a discussion of the theoretical bases for and the graphs and the properties of the large number of

functions dealt with.

Vol. I, Parts III, IV, and V discuss, give tables and bibliography covering a variety of interpolation procedures. Five tables of log Γ and one of $1/\Gamma(re^{\theta i})$ are given. The argument interval is small and the number of decimal places varies from 10 to 12 to 15. Six tables give or involve the Psi function, which is the digamma function as defined by Eleanor Pairman, in *Tracts for Computers*, No. 1, 1919. The argument interval is small and the number of decimal places varies from 10 to 12 to 15 to 16 to 20.

Vol. II treats of: (*a*) The polygamma functions, giving tables of trigamma, tetragamma, pentagamma, and hexagamma functions, employing a fairly small interval in the argument and entries from 10 up to 20 decimal places. (*b*) Bernoulli polynomials and numbers. (*c*) Euler polynomials and numbers. (*d*) Gram polynomials. (*e*) functions of polynomial approximation.

1934-1938. BIOMETRIKA PUBLICATIONS. Karl Pearson, Ed. University College, London, W. C. 1.

Tables of the Incomplete Beta-Function. 1934. The complete beta-function is defined.

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$$

and the incomplete beta-function

$$B_x(p,q) = \int_0^x x^{p-1}(1-x)^{q-1}\,dx$$

The magnitude tabled is

$$I_x(p,q) = B_x(p,q)/B(p,q)$$

This trivariate tabling required changes in interval in p and q, but a uniform interval in x of .01 for the entire range $x = 0$ to $x = 1$ was employed. Even so the tables require 494 pages. We note the relationship between $I_x(p,q)$ and the variance ratio F_{ij} (see [9:21]): i, j, F_{ij}, and

P_{ij} refer to Kelley's notation and p, q, and x to Pearson's.

$$i = 2q;\; j=2p;\; \frac{j}{iF_{ij}+j} = x;\; P_{ij} = I_x(p,q)$$

It may be stated that P_{ij} via these Pearson Tables is usually one and sometimes two decimal places more accurate than by the normalizing transformation of Chapter IX, Section 5 herein.

Tables of the Incomplete Gamma Function. Reissue, 1934.

$$I\,(u,p) = \left(\int_0^{u\sqrt{p+1}} \epsilon^{-v}\, v^p\, dv\right) \Big/\ \Gamma\,(p+1)$$

is tabled. It is the probability integral of a Type III distribution. [7 decimal places in Tables I and II and 8 in Table III.] Table I arguments p = .0 (.1) 5.0 (.2) 50.0 and u = .0 (.1) 16.9. Table II arguments p = −1.00 (.05) .00 and u = .0 (.1) 48.0. Table III is of log $i'(u,p)$, wherein $I'(u,p) = I(u,p)/u^{p+1}$. Arguments u = .0 (.1) 1.5 and p = −1.00 (.05) .0 (.1) 10.0. Table IV provides "constants of the skew curve $y = y_0 x^p e^{-x}$" and Table V gives "five figure values of $I(u,p)$."

(F. N. David) Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples. 1938. Gives the distribution of r for ρ (Kelley's $\tilde{r}$)=.0 (.1) .9, and N = 3 (1) 25, 50, 100, 200, 400.

1934. H. B. Dwight. TABLES OF INTEGRALS AND OTHER MATHEMATICAL DATA, 222 pages. Similar to, but somewhat more extensive then, Peirce's Table.

1935. Truman L. Kelley. ESSENTIAL TRAITS OF MENTAL LIFE. Table XXII of Trigonometric functions used in making rotations of axes. Tabled, to 7 decimal places or 7 significant figures, are tan 2θ, $\cos\theta$, $\sin\theta$, $\cos^2\theta$, $\sin^2\theta$, $2\sin\theta\,\cos\theta$,

$\sin\theta \cos\theta$, for the argument $\theta = 0°00'$ (1') $1°00'$ (3') $2°0'$ (6') $3°0'$ (10') $45°0'$.

1938. R. A. Fisher and F. Yates. STATISTICAL TABLES FOR BIOLOGICAL, AGRICULTURAL, AND MEDICAL RESEARCH.

Table III [to 3 decimals] *Distribution of t* [Student's t]. Argument is n, the d. o. f., for $n = 1$ (1) 30, 40, 60, 120, ∞.

Table IV [4 and 5 figures] *Distribution of* χ^2. Argument is n, the d. o. f., for $n = 1$ (1) 30.

Table V (C. G. Colcord, L. S. Deming, R. A. Fisher, H. W. Norton and Y. Yates) [4 decimal places] *Distribution of z.* "z may be defined as the difference of one-half the natural logarithms of two different estimates of variance, one based on n_1 and the other on n_2 degrees of freedom." Arguments are $n_1 = 1$ (1) 6, 8, 12, 24, ∞, and $n_2 = 1$ (1) 30, 40, 60, 120, ∞. Tables for percentage points 20, 5, 1, and .1 are given. The second part of Table V tables [generally 3 figures] the variance ratio instead of z for the same arguments. Cf. Merrington and Thompson, "Tables of percentage points of the inverted beta (F) distribution," and Thompson, "Tables of percentage points of the incomplete beta-function," and also Comrie and Hartley, "Table of Lagrangian coefficients for harmonic interpolation in certain tables of percentage points."

Table VII [4 and 5 places] *Transformation of r to z.* Here z is defined as in [10:43]. The argument is $z = .0$ (.1) 3, 4.

Table XV Latin squares. Sets up to 9X9 are given.

Table XX Scores for ordinal (or ranked) data. Herein is given "the average deviate of the rth largest of N observations drawn from a normal distribution having unit variance." Sample sizes up to $N = 40$.

Table XXIII Orthogonal polynomials. (R. A. Fisher and M. F. Yates and V. Satakopan) From

$n = 3$ to $n = 51$.

Table XXVI [5 decimal places] "Natural logarithms." Arguments 1.00 (.01) 10. (1) 999.

Table XXVII Squares. Arguments 1 (1) 999.

Table XXVIII [5 figures] *Square roots.* Arguments 100 (1) 999, and also 10.0 (.1) 99.9.

Table XXIX [6 places] *Reciprocals.* Arguments 1.00 (.01) 9.99.

Table XXX [6 figures in the factorial and 7 decimal places in the logarithm of the factorial] *Factorials.* Arguments 1 (1) 200.

Table XXXIII [2 figures] *Random numbers.* Six sets of 1250 each of two digit random numbers.

1939 to 1944. NATIONAL BUREAU OF STANDARDS TABLES,- prepared by the **FEDERAL WORKS AGENCY,** Works Project Administration, for the City of New York, **Arnold N. Lowan, Technical Director.**

Table of Natural Logarithms, [16 decimal places] Vol. I Integers from 1 to 50,000. Vol. II Integers from 50,000 to 100,000. Vol. III Decimal numbers from 0 to 5 by intervals of .0001. Vol. IV Decimal numbers from 5 to 10 by intervals of .0001. 1941.

Tables of the exponential function e^x. [12 to 18 decimals] Argument $x = -2.5000$, (.0001) 2.500 (.001) 5.00 (.01) 10.00. 1939.

Tables of circular and hyperbolic sines and cosines. [9 decimal places] Argument x (in radians) from 0 to 2 by intervals of .0001. 1939.

Tables of sines and cosines. [8 decimal places] Argument x (in radians) from 0 to 25 by intervals of .001. 1940.

Tables of probability functions. [15 decimal places]

Vol. I, 1941.

$$\frac{2}{\sqrt{\pi}} e^{-x^2}$$ (= $\sqrt{8}z$ of Table XV C herein)

$$\frac{2}{\sqrt{\pi}} \int_0^x e^{-\alpha^2}\, d\alpha$$ (=$1-2q$ = $p-q$ of Table XV C herein)

Argument x (= $x/\sqrt{2}$ of Table XV C herein) = .0000 (.0001) 1.000 (.001) 5.946.

Vol. II, 1942.

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad (= z \text{ of Table XV C herein})$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{\frac{-\alpha^2}{2}} \, d\alpha (= 1-2q = p-q \text{ of Table XV C herein})$$

Argument x (= x of Table XV C herein) = .0000 (.0001) 1.000 (.001) 8.285.

(*Herbert E. Salzer*) *Table of coefficients for inverse interpolation with central differences.* [10 places] $m = (y-y_0)/(y_1-y_0)$. Argument m = .000 (.001) .500. First printed in Journal of Mathematics and Physics, Vol. 22, no. 4, Dec., 1943.

(*Herbert E. Salzer*) *Table of coefficients for inverse interpolation with advancing differences.* [10 places] m as in preceding. First printed in Journal of Mathematics and Physics, Vol. 23, no. 2, May, 1944.

Tables of Lagrangian Interpolation Coefficients, 1944. n is the number of points used in a table having equally spaced arguments and p is the same as herein [13:93]. The tables give exact values, or values to 10 decimal places.

For n = 3 the argument p = −1 (.0001) 1
For n = 4 the argument p = −1 (.001) 0 (.0001) 1 (.001) 2
For n = 5 the argument p = −2 (.001) 2
For n = 6 the argument p = −2 (.01) 0 (.001) 1 (.01) 3
For n = 7 the argument p = −3 (.01) −1 (.001) 1 (.01) 3
For n = 8 the argument p = −3 (.1) 0 (.001) 1 (.1) 4
For n = 9 the argument p = −4 (.1) 4

For $n = 10$ the argument $p = -4\ (.1)\ 5$

For $n = 11$ the argument $p = -5\ (.1)\ 5$

Certain supplementary tables are given. These main tables are more extensive than Kelley's Lagrangian interpolation coefficient tables, especially in that they include 5, 7, 9, and 11 point interpolation, in that the interval in p for 4 point is .0001 instead of .001, for 6 point is .001 instead of .01, for 8 point is .001 instead of .1, and in that 9 place decimal or exact values are given for $n = 3$ whereas Kelley rounded off to 5 decimal places to simplify use in the usual situation wherein quadric interpolation will not yield accuracy beyond an added 4 or 5 figures even though coefficient values are exact.

1939. William Fleetwood Sheppard. THE PROBABILITY INTEGRAL. Completed and edited by the British Association for the Advancement of Science, Committee for the Calculation of Mathematical Tables. Cambridge University Press. Table I gives $(F)x$ (= Kelley's q_x/z_x, — [8:02] and [8:03] to 12 decimals for argument $x = .00\ (.01)\ 10.00$. Table II gives $F(x)$ to 24 decimals for $x = .0\ (.1)\ 10.0$. Table IV gives $L(x)$ (= $-\ell n\ q_x$) to 16 decimals for $x = .0\ (.1)\ 10.0$. Table V gives $\ell(x)$ (= $\log q_x$) to 12 decimals for $x = .0\ (.1)\ 10.0$. Table VI gives $\ell(x)$ to 8 decimals for $x = .00\ (.01)\ 10.00$.

1941. L. J. Comrie and H. O. Hartley. *Table of Lagrangian coefficients for harmonic interpolation in certain tables of percentage points*, Biom., Vol. 32, 1941, pp. 183-186.

1941. Catherine M. Thompson. *Table of percentage of the χ^2 distribution*, Biom., Vol. 32, 1941, pp. 187-191. χ^2, to 6 significant figures, is given for P = .995, .99, .975, .95, .90, .75, .50, .25, .10, .05, .025, .01, .005, and for d.o.f. = 1 (1) 30 (10) 100.

1941. Catherine M. Thompson. *Tables of per-*

centage points of the incomplete beta function. Biom., Vol. 32, pp. 168-181. The function $I_x(p,q)$ is that defined by Pearson in *Tables of the incomplete beta-function.* The percentage points are the values of x which, for given values of P, p, and q, satisfy the equation $I_x(p,q) = P$. x, to 5 significant figures, is given for the following values of P, of $2p$ [= Kelley's d.o.f. i] and of $2q$ [= Kelley's d.o.f. j]: P = .50, .25, .10, .05, .025, .01, .005, $2q$ = 1 (1) 10, 12, 15, 20, 24, 30, 40, 60, 120, ∞. $2p$ = 1 (1) 30, 40, 60, 120, ∞. Note that Fisher and Yates' Table V (by H. W. Norton) giving percentage points for the related function "z" for P = .20 and .001, supplement this table. Also cf. Comrie and Hartley, *Table of Lagrangian coefficients for harmonic interpolation in certain tables of percentage points.*

1942. Edward Charles Dixon Molina. **POISSON'S EXPONENTIAL BINOMIAL LIMIT.** Table I,—individual terms, and Table II,—cumulated terms. [6 decimal places] Argument interval varies from .001 to 1.

1942. **SMITHSONIAN MATHEMATICAL TABLES,** *Hyperbolic Functions.* Fifth reprint, 1942, by *George F. Becker and C. E. Van Orstand.* Derivatives recorded in addition to items noted below.

Table I. [5 decimal places] *Logarithms of hyperbolic functions.* Argument u = .0000 (.0001) .100 (.001) 3.00, (.01) 6.00.

Table II. [5 decimal places] *Natural hyperbolic functions.* Same arguments as in Table I. As an illustration of the use of this table, note that u herein is Fisher's z from r and *tanh* is r.

Table III. [5 decimal places] *Natural and logarithmic circular functions.* Argument u = .0000 (.0001) .100 (.001) 1.600. A supplementary table using a coarser argument provides the functions for higher values of u.

Table IV. Ascending and descending exponen-

tial and $\log_{10}(e^u)$. For argument u is given $\log_{10}(e^u)$ to 7 decimal places; e^u to 7 or more significant figures; e^{-u} to 7 decimal places. Argument u = .000 (.001) 3.00 (.01) . . . 15.00. A supplementary table using a coarser argument provides the functions for higher values of u.

Table V. [6 places] *Natural logarithms.* Argument u = 0 (1) 1000 and thence by irregular arguments to 10,000. The same table serviceable to obtain e^x and e^{-x} for irregular arguments in x.

Table VI. [7 decimal places] *The Gudermannian.* Argument u = .000 (.001) 3.00 (.01) 6.00. Table VII gives the "Anti-Gudermannian."

1942-1943. *Egon S. Pearson and H. O. Hartley. The probability integral of the range in samples of N observations from a normal population.* **Biom.**, Vol. 32, 1942, pp. 301-310. The arguments are N = 2 (1) 20, and W, the ratio of a specific range to the population standard deviation, = .00 (.05) 7.25. The tabled entry is, to 4 decimal places, the cumulated frequency to the (N, W) point in question. The same authors in *Tables of the probability integral of the studentized range,* **Biom.**, Vol. 32, 1943, pp. 89-99, employ a second sample to secure an independent estimate of the population standard deviation, and thence tables for calculating the desired probability integral.

1943. **M. Merrington and Catherine M. Thompson.** *Tables of percentage points of the inverted beta (F) distribution.* **Biom.**, Vol. 33, 1943, pp. 73-88. F is Snedecor's F (Kelley's F_{ij}), the variance ratio. F is tabled, to 5 significant figures, for P = .50, .25, .10, .05, .025, .01, .005, for ν_1 (Kelley's d.o.f. i) = 1 (1) 10, 12, 15, 24, 30, 40, 60, 120, ∞, and for ν_2 (Kelley's d.o.f. j) = 1 (1) 30, 40, 60, 120, ∞. Note that Fisher and Yates' Table V (by H. W. Norton), giving percentage points for P = .20 and .001,

supplement this table. Also cf. Comrie and Hartley, *Table of Lagrangian coefficients for harmonic interpolation in certain tables of percentage points.*

1944. BESSEL FUNCTIONS. For references to the extensive tables of these functions and to literature about them see *Harry Bateman and Raymond Claire Archibald, A guide to Tables of Bessel Functions,* Mathematical Tables and Other Aids to Computation, Vol. I, No. 7, July, 1944.

In press, 1947. The Truman Lee KELLEY STATISTICAL TABLES.

Table I. Eight place normal distribution, simple correlation, and probability functions. The argument p = .5000 (.0001) .9999, or its complement, $q = 1-p$, may variously be considered the proportionate area in the tail of a normal distribution, a probability, or a positive correlation coefficient, a sine, or a cosine of an angle, etc. Tabled are x and z, the deviate and ordinate in a unit normal distribution, $\sqrt{pq}$, $\sqrt{1-p^2}$, $\sqrt{1-q^2}$, and also E^{II} and E^{III}, the errors in linear and in quadric interpolation, and, for p and q, E^{-II} and E^{-III}, the errors in inverse linear and quadric interpolation. A supplementary table for $p > .9999$ is given. Table XV C herein is an abridgement and modification of this table.

Table II. Five place [fifth place rounded off] *three-point interpolation coefficients.* Argument p (a proportionate distance between two tabled arguments in any table having equally spaced arguments) = .0000 (.0001) .5000. Table XV A herein is an abridgement of this table.

Table III. Seven place [seventh place compensatorially rounded off] *four-point interpolation coefficients.* Argument p = .000 (.001) .500. Table XV B herein is an abridgement of this table.

Table IV. Ten place [tenth place rounded off]

six-point interpolation coefficients. Argument p = .00 (.01) .50.

Table V. Eleven place (exact) *eight point interpolation coefficients.* Argument p=.0 (.1) 1.0.

Table VI. Four-place χ^2 functions. Argument $\chi^2 / \sqrt{\text{d.o.f.}}$ = .0 (.1) 4.1. P, the probability that a divergence as great as χ^2 will arise as a matter of chance, is tabled for 1 (1) 10, 12, 15, 19, 24, 30 d.o.f. Also direct and inverse interpolation errors E^{ii}, E^{iii}, E^{-ii}, E^{-iii}.

Table VII. Eight place square roots, cube roots (for all three place numbers) and natural logarithms, for N = 1.0 (.01) 10.0. Table XV D herein is an abridgement of this table.

Table VIII. Eight place θ_1 functions for normalizing F_{ij}, the variance ratio. Argument i (= d.o.f.) = 1 (1) 100 (10) 200 (20) 300 (50) 500, (100) 1000, ∞. Table IX E herein is an abridgement of this table.

SECTION 2. LAGRANGIAN INTERPOLATION COEFFICIENTS

Table XV A, which is a 90 per cent abridgement of the three-point Lagrangian interpolation coefficients given in *The Kelley Statistical Tables* (in press, 1947) gives, for p = .001 (.001) .500, the c_0, c_1, and c_2 coefficients of [13:98a], with such compensatory rounding off to five decimal places as to make the rounding off error very small. (See Kelley, *Tables*, in press 1947) When p <.5 coefficients and tabled entries t_0, t_1, and t_2 are multiplied and summed as in [13:98a], and when p > 5 then q is the desired fraction and computation proceeds in the reverse direction using t_1, t_0, and t_{-1}.

Table XV B, a 90 per cent abridgement of the four-point coefficients given in *The Kelley Statistical Tables,* gives, for p = .00 (.01) .50, the c_{-1}, c_0, c_1, c_2 coefficients of [13:99a].

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.000	.00000	1.00000	.00000	.030	.01455	.99910	.01545
.001	.00050	1.00000	.00050	.031	.01502	.99904	.01598
.002	.00100	1.00000	.00100	.032	.01549	.99898	.01651
.003	.00150	1.00000	.00150	.033	.01596	.99892	.01704
.004	.00199	.99998	.00201	.034	.01642	.99884	.01758
.005	.00249	.99998	.00251	.035	.01689	.99878	.01811
.006	.00298	.99996	.00302	.036	.01735	.99870	.01865
.007	.00348	.99996	.00352	.037	.01782	.99864	.01918
.008	.00397	.99994	.00403	.038	.01828	.99856	.01972
.009	.00446	.99992	.00454	.039	.01874	.99848	.02026
.010	.00495	.99990	.00505	.040	.01920	.99840	.02080
.011	.00544	.99988	.00556	.041	.01966	.99832	.02134
.012	.00593	.99986	.00607	.042	.02012	.99824	.02188
.013	.00642	.99984	.00658	.043	.02058	.99816	.02242
.014	.00690	.99980	.00710	.044	.02103	.99806	.02297
.015	.00739	.99978	.00761	.045	.02149	.99798	.02351
.016	.00787	.99974	.00813	.046	.02194	.99788	.02406
.017	.00836	.99972	.00864	.047	.02240	.99780	.02460
.018	.00884	.99968	.00916	.048	.02285	.99770	.02515
.019	.00932	.99964	.00968	.049	.02330	.99760	.02570
.020	.00980	.99960	.01020	.050	.02375	.99750	.02625
.021	.01028	.99956	.01072	.051	.02420	.99740	.02680
.022	.01076	.99952	.01124	.052	.02465	.99730	.02735
.023	.01124	.99948	.01176	.053	.02510	.99720	.02790
.024	.01171	.99942	.01229	.054	.02554	.99708	.02846
.025	.01219	.99938	.01281	.055	.02599	.99698	.02901
.026	.01266	.99932	.01334	.056	.02643	.99686	.02957
.027	.01314	.99928	.01386	.057	.02688	.99676	.03012
.028	.01361	.99922	.01439	.058	.02732	.99664	.03068
.029	.01408	.99916	.01492	.059	.02776	.99652	.03124
.030	.01455	.99910	.01545	.060	.02820	.99640	.03180

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.060	.02820	.99640	.03180	.090	.04095	.99190	.04905
.061	.02864	.99628	.03236	.091	.04136	.99172	.04964
.062	.02908	.99616	.03292	.092	.04177	.99154	.05023
.063	.02952	.99604	.03348	.093	.04218	.99136	.05082
.064	.02995	.99590	.03405	.094	.04258	.99116	.05142
.065	.03039	.99578	.03461	.095	.04299	.99098	.05201
.066	.03082	.99564	.03518	.096	.04339	.99078	.05261
.067	.03126	.99552	.03574	.097	.04380	.99060	.05320
.068	.03169	.99538	.03631	.098	.04420	.99040	.05380
.069	.03212	.99524	.03688	.099	.04460	.99020	.05440
.070	.03255	.99510	.03745	.100	.04500	.99000	.05500
.071	.03298	.99496	.03802	.101	.04540	.98980	.05560
.072	.03341	.99482	.03859	.102	.04580	.98960	.05620
.073	.03384	.99468	.03916	.103	.04620	.98940	.05680
.074	.03426	.99452	.03974	.104	.04659	.98918	.05741
.075	.03469	.99438	.04031	.105	.04699	.98898	.05801
.076	.03511	.99422	.04089	.106	.04738	.98876	.05862
.077	.03554	.99408	.04146	.107	.04778	.98856	.05922
.078	.03596	.99392	.04204	.108	.04817	.98834	.05983
.079	.03638	.99376	.04262	.109	.04856	.98812	.06044
.080	.03680	.99360	.04320	.110	.04895	.98790	.06105
.081	.03722	.99344	.04378	.111	.04934	.98768	.06166
.082	.03764	.99328	.04436	.112	.04973	.98746	.06227
.083	.03806	.99312	.04494	.113	.05012	.98724	.06288
.084	.03847	.99294	.04553	.114	.05050	.98700	.06350
.085	.03889	.99278	.04611	.115	.05089	.98678	.06411
.086	.03930	.99260	.04670	.116	.05127	.98654	.06473
.087	.03972	.99244	.04728	.117	.05166	.98632	.06534
.088	.04013	.99226	.04787	.118	.05204	.98608	.06596
.089	.04054	.99208	.04846	.119	.05242	.98584	.06658
.090	.04095	.99190	.04905	.120	.05280	.98560	.06720

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.120	.05280	.98560	.06720	.150	.06375	.97750	.08625
.121	.05318	.98536	.06782	.151	.06410	.97720	.08690
.122	.05356	.98512	.06844	.152	.06445	.97690	.08755
.123	.05394	.98488	.06906	.153	.06480	.97660	.08820
.124	.05431	.98462	.06969	.154	.06514	.97628	.08886
.125	.05469	.98438	.07031	.155	.06549	.97598	.08951
.126	.05506	.98412	.07094	.156	.06583	.97566	.09017
.127	.05544	.98388	.07156	.157	.06618	.97536	.09082
.128	.05581	.98362	.07219	.158	.06652	.97504	.09148
.129	.05618	.98336	.07282	.159	.06686	.97472	.09214
.130	.05655	.98310	.07345	.160	.06720	.97440	.09280
.131	.05692	.98284	.07408	.161	.06754	.97408	.09346
.132	.05729	.98258	.07471	.162	.06788	.97376	.09412
.133	.05766	.98232	.07534	.163	.06822	.97344	.09478
.134	.05802	.98204	.07598	.164	.06855	.97310	.09545
.135	.05839	.98178	.07661	.165	.06889	.97278	.09611
.136	.05875	.98150	.07725	.166	.06922	.97244	.09678
.137	.05912	.98124	.07788	.167	.06956	.97212	.09744
.138	.05948	.98096	.07852	.168	.06989	.97178	.09811
.139	.05984	.98068	.07916	.169	.07022	.97144	.09898
.140	.06020	.98040	.07980	.170	.07055	.97110	.09945
.141	.06056	.98012	.08044	.171	.07088	.97076	.10012
.142	.06092	.97984	.08108	.172	.07121	.97042	.10079
.143	.06128	.97956	.08172	.173	.07154	.97008	.10146
.144	.06163	.97926	.08237	.174	.07186	.96972	.10214
.145	.06199	.97898	.08301	.175	.07219	.96938	.10281
.146	.06234	.97868	.08366	.176	.07251	.96902	.10349
.147	.06270	.97840	.08430	.177	.07284	.96868	.10416
.148	.06305	.97810	.08495	.178	.07316	.96832	.10484
.149	.06340	.97780	.08560	.179	.07348	.96796	.10552
.150	.06375	.97750	.08625	.180	.07380	.96760	.10620

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.180	.07380	.96760	.10620	.210	.08295	.95590	.12705
.181	.07412	.96724	.10688	.211	.08324	.95548	.12776
.182	.07444	.96688	.10756	.212	.08353	.95506	.12847
.183	.07476	.96652	.10824	.213	.08382	.95464	.12918
.184	.07507	.96614	.10893	.214	.08410	.95420	.12990
.185	.07539	.96578	.10961	.215	.08439	.95378	.13061
.186	.07570	.96540	.11030	.216	.08467	.95334	.13133
.187	.07602	.96504	.11098	.217	.08496	.95292	.13204
.188	.07633	.96466	.11167	.218	.08524	.95248	.13276
.189	.07664	.96428	.11236	.219	.08552	.95204	.13348
.190	.07695	.96390	.11305	.220	.08580	.95160	.13420
.191	.07726	.96352	.11374	.221	.08608	.95116	.13492
.192	.07757	.96314	.11443	.222	.08636	.95072	.13564
.193	.07788	.96276	.11512	.223	.08664	.95028	.13636
.194	.07818	.96236	.11582	.224	.08691	.94982	.13709
.195	.07849	.96198	.11651	.225	.08719	.94938	.13781
.196	.07879	.96158	.11721	.226	.08746	.94892	.13854
.197	.07910	.96120	.11790	.227	.08774	.94848	.13926
.198	.07940	.96080	.11860	.228	.08801	.94802	.13999
.199	.07970	.96040	.11930	.229	.08828	.94756	.14072
.200	.08000	.96000	.12000	.230	.08855	.94710	.14145
.201	.08030	.95960	.12070	.231	.08812	.94664	.14218
.202	.08060	.95920	.12140	.232	.08909	.94618	.14291
.203	.08090	.95880	.12210	.233	.08936	.94572	.14364
.204	.08119	.95838	.12281	.234	.08962	.94524	.14438
.205	.08149	.95798	.12351	.235	.08989	.94478	.14511
.206	.08178	.95756	.12422	.236	.09015	.94430	.14585
.207	.08208	.95716	.12492	.237	.09042	.94384	.14658
.208	.08237	.95674	.12563	.238	.09068	.94336	.14732
.209	.08266	.95632	.12634	.239	.09094	.94288	.14806
.210	.08295	.95590	.12705	.240	.09120	.94240	.14880

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.240	.09120	.94240	.14880	.270	.09855	.92710	.17145
.241	.09146	.94192	.14954	.271	.09878	.92656	.17222
.242	.09172	.94144	.15028	.272	.09901	.92602	.17299
.243	.09198	.94096	.15102	.273	.09924	.92548	.17376
.244	.09223	.94046	.15177	.274	.09946	.92492	.17454
.245	.09249	.93998	.15251	.275	.09969	.92438	.17531
.246	.09274	.93948	.15326	.276	.09991	.92382	.17609
.247	.09300	.93900	.15400	.277	.10014	.92328	.17686
.248	.09325	.93850	.15475	.278	.10036	.92272	.17764
.249	.09350	.93800	.15550	.279	.10058	.92216	.17842
.250	.09375	.93750	.15625	.280	.10080	.92160	.17920
.251	.09400	.93700	.15700	.281	.10102	.92104	.17998
.252	.09425	.93650	.15775	.282	.10124	.92048	.18075
.253	.09450	.93600	.15850	.283	.10146	.91992	.18154
.254	.09474	.93548	.15926	.284	.10167	.91934	.18233
.255	.09499	.93498	.16001	.285	.10189	.91878	.18311
.256	.09523	.93446	.16077	.286	.10210	.91820	.18390
.257	.09548	.93396	.16152	.287	.10232	.91764	.18468
.258	.09572	.93344	.16228	.288	.10253	.91706	.18547
.259	.09596	.93292	.16304	.289	.10274	.91648	.18626
.260	.09620	.93240	.16380	.290	.10295	.91590	.18705
.261	.09644	.93188	.16456	.291	.10316	.91532	.18784
.262	.09668	.93136	.16532	.292	.10337	.91474	.18863
.263	.09692	.93084	.16608	.293	.10358	.91416	.18942
.264	.09715	.93030	.16685	.294	.10378	.91356	.19022
.265	.09739	.92978	.16761	.295	.10399	.91298	.19101
.266	.09762	.92924	.16838	.296	.10419	.91238	.19181
.267	.09786	.92872	.16914	.297	.10440	.91180	.19260
.268	.09809	.92818	.16991	.298	.10460	.91120	.19340
.269	.09832	.92764	.17068	.299	.10480	.91060	.19420
.270	.09855	.92710	.17145	.300	.10500	.91000	.19500

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.300	.10500	.91000	.19500	.330	.11055	.89110	.21945
.301	.10520	.90940	.19580	.331	.11072	.89044	.22028
.302	.10540	.90880	.19660	.332	.11089	.88978	.22111
.303	.10560	.90820	.19740	.333	.11106	.88912	.22194
.304	.10579	.90758	.19821	.334	.11122	.88844	.22278
.305	.10599	.90698	.19901	.335	.11139	.88778	.22361
.306	.10618	.90636	.19982	.336	.11155	.88710	.22445
.307	.10638	.90576	.20062	.337	.11172	.88644	.22528
.308	.10657	.90514	.20143	.338	.11188	.88576	.22612
.309	.10676	.90452	.20224	.339	.11204	.88508	.22696
.310	.10695	.90390	.20305	.340	.11220	.88440	.22780
.311	.10714	.90328	.20386	.341	.11236	.88372	.22864
.312	.10733	.90266	.20467	.342	.11252	.88304	.22948
.313	.10752	.90204	.20548	.343	.11268	.88236	.23032
.314	.10770	.90140	.20630	.344	.11283	.88166	.23117
.315	.10789	.90078	.20711	.345	.11299	.88098	.23201
.316	.10807	.90014	.20793	.346	.11314	.88028	.23286
.317	.10826	.89952	.20874	.347	.11330	.87960	.23370
.318	.10844	.89888	.20956	.348	.11345	.87890	.23455
.319	.10862	.89824	.21038	.349	.11360	.87820	.23540
.320	.10880	.89760	.21120	.350	.11375	.87750	.23625
.321	.10898	.89696	.21202	.351	.11390	.87680	.23710
.322	.10916	.89632	.21284	.352	.11405	.87610	.23795
.323	.10934	.89568	.21366	.353	.11420	.87540	.23880
.324	.10951	.89502	.21449	.354	.11434	.87468	.23966
.325	.10969	.89438	.21531	.355	.11449	.87398	.24051
.326	.10986	.89372	.21614	.356	.11463	.87326	.24137
.327	.11004	.89308	.21696	.357	.11478	.87256	.24222
.328	.11021	.89242	.21779	.358	.11492	.87184	.24308
.329	.11038	.89176	.21862	.359	.11506	.87112	.24394
.330	.11055	.89110	.21945	.360	.11520	.87040	.24480

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.360	.11520	.87040	.24480	.390	.11895	.84790	.27105
.361	.11534	.86968	.24566	.391	.11906	.84712	.27194
.362	.11548	.86896	.24652	.392	.11917	.84634	.27283
.363	.11562	.86824	.24738	.393	.11928	.84556	.27372
.364	.11575	.86750	.24825	.394	.11938	.84476	.27462
.365	.11589	.86678	.24911	.395	.11949	.84398	.27551
.366	.11602	.86604	.24998	.396	.11959	.84318	.27641
.367	.11616	.86532	.25084	.397	.11970	.84240	.27730
.368	.11629	.86458	.25171	.398	.11980	.84160	.27820
.369	.11642	.86384	.25258	.399	.11990	.84080	.27910
.370	.11655	.86310	.25345	.400	.12000	.84000	.28000
.371	.11668	.86236	.25432	.401	.12010	.83920	.28090
.372	.11681	.86162	.25519	.402	.12020	.83840	.28180
.373	.11694	.86088	.25606	.403	.12030	.83760	.28270
.374	.11706	.86012	.25694	.404	.12039	.83678	.28361
.375	.11719	.85938	.25781	.405	.12049	.83598	.28451
.376	.11731	.85862	.25869	.406	.12058	.83516	.28542
.377	.11744	.85788	.25956	.407	.12068	.83436	.28632
.378	.11756	.85712	.26044	.408	.12077	.83354	.28723
.379	.11768	.85636	.26132	.409	.12086	.83272	.28814
.380	.11780	.85560	.26220	.410	.12095	.83190	.28905
.381	.11792	.85484	.26308	.411	.12104	.83108	.28996
.382	.11804	.85408	.26396	.412	.12113	.83026	.29087
.383	.11816	.85332	.26484	.413	.12122	.82944	.29178
.384	.11827	.85254	.26573	.414	.12130	.82860	.29270
.385	.11839	.85178	.26661	.415	.12139	.82778	.29361
.386	.11850	.85100	.26750	.416	.12147	.82694	.29453
.387	.11862	.85024	.26838	.417	.12156	.82612	.29544
.388	.11873	.84946	.26927	.418	.12164	.82528	.29636
.389	.11884	.84868	.27016	.419	.12172	.82444	.29728
.390	.11895	.84790	.27105	.420	.12180	.82360	.29820

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.420	.12180	.82360	.29820	.450	.12375	.79750	.32625
.421	.12188	.82276	.29912	.451	.12380	.79660	.32720
.422	.12196	.82192	.30004	.452	.12385	.79570	.32815
.423	.12204	.82108	.30096	.453	.12390	.79480	.32910
.424	.12211	.82022	.30189	.454	.12394	.79388	.33006
.425	.12219	.81938	.30281	.455	.12399	.79298	.33101
.426	.12226	.81852	.30374	.456	.12403	.79206	.33197
.427	.12234	.81768	.30466	.457	.12408	.79116	.33292
.428	.12241	.81682	.30559	.458	.12412	.79024	.33388
.429	.12248	.81596	.30652	.459	.12416	.78932	.33484
.430	.12255	.81510	.30745	.460	.12420	.78840	.33580
.431	.12262	.81424	.30838	.461	.12424	.78748	.33676
.432	.12269	.81338	.30931	.462	.12428	.78656	.33772
.433	.12276	.81252	.31024	.463	.12432	.78564	.33868
.434	.12282	.81164	.31118	.464	.12435	.78470	.33965
.435	.12289	.81078	.31211	.465	.12439	.78378	.34061
.436	.12295	.80990	.31305	.466	.12442	.78284	.34158
.437	.12302	.80904	.31398	.467	.12446	.78192	.34254
.438	.12308	.80816	.31492	.468	.12449	.78098	.34351
.439	.12314	.80728	.31586	.469	.12452	.78004	.34448
.440	.12320	.80640	.31680	.470	.12455	.77910	.34545
.441	.12326	.80552	.31774	.471	.12458	.77816	.34642
.442	.12332	.80464	.31868	.472	.12461	.77722	.34739
.443	.12338	.80376	.31962	.473	.12464	.77628	.34836
.444	.12343	.80286	.32057	.474	.12466	.77532	.34934
.445	.12349	.80198	.32151	.475	.12469	.77438	.35031
.446	.12354	.80108	.32246	.476	.12471	.77342	.35129
.447	.12360	.80020	.32340	.477	.12474	.77248	.35226
.448	.12365	.79930	.32435	.478	.12476	.77152	.35324
.449	.12370	.79840	.32530	.479	.12478	.77056	.35422
.450	.12375	.79750	.32625	.480	.12480	.76960	.35520

TABLE XV A

THREE-POINT INTERPOLATION COEFFICIENTS

p	c_0 −	c_1 +	c_2 +	p	c_0 −	c_1 +	c_2 +
.480	.12480	.76960	.35520	.490	.12495	.75990	.36505
.481	.12482	.76864	.35618	.491	.12496	.75892	.36604
.482	.12484	.76768	.35716	.492	.12497	.75794	.36703
.483	.12486	.76672	.35814	.493	.12498	.75696	.36802
.484	.12487	.76574	.35913	.494	.12498	.75596	.36902
.485	.12489	.76478	.36011	.496	.12499	.75498	.37001
.486	.12490	.76380	.36110	.496	.12499	.75398	.37101
.487	.12492	.76284	.36208	.497	.12500	.75300	.37200
.488	.12493	.76186	.36307	.498	.12500	.75200	.37300
.489	.12494	.76088	.36406	.499	.12500	.75100	.37400
.490	.12495	.75990	.36505	.500	.12500	.75000	.37500

TABLE XV B

FOUR-POINT INTERPOLATION COEFFICIENTS

p ($p<.5$)	c_{-1} −	c_0 +	c_1 +	c_2 −	p ($p>.5$)
.00	.00000 00	1.00000 00	.00000 00	.00000 00	1.00
.01	.00328 35	.99490 05	.01004 95	.00166 65	.99
.02	.00646 80	.98960 40	.02019 60	.00333 20	.98
.03	.00955 45	.98411 35	.03043 65	.00499 55	.97
.04	.01254 40	.97843 20	.04076 80	.00665 60	.96
.05	.01543 75	.97256 25	.05118 75	.00831 25	.95
.06	.01823 60	.96650 80	.06169 20	.00996 40	.94
.07	.02094 05	.96027 15	.07227 85	.01160 95	.93
.08	.02355 20	.95385 60	.08294 40	.01324 80	.92
.09	.02607 15	.94726 45	.09363 55	.01487 85	.91
.10	.02850 00	.94050 00	.10450 00	.01650 00	.90
.11	.03083 85	.93356 55	.11538 45	.01811 15	.89
.12	.03308 80	.92646 40	.12633 60	.01971 20	.88
.13	.03524 95	.91919 85	.13735 15	.02130 05	.87
.14	.03732 40	.91177 20	.14842 80	.02287 60	.86
.15	.03931 25	.90418 75	.15956 25	.02443 75	.85
.16	.04121 60	.89644 80	.17075 20	.02598 40	.84
.17	.04303 55	.88855 65	.18199 35	.02751 45	.83
.18	.04477 20	.88051 60	.19328 40	.02902 80	.82
.19	.04642 65	.87232 95	.20462 05	.03052 35	.81
.20	.04800 00	.86400 00	.21600 00	.03200 00	.80
.21	.04949 35	.85553 05	.22741 95	.03345 65	.79
.22	.05090 80	.84692 40	.23887 60	.03489 20	.78
.23	.05224 45	.83818 35	.25036 65	.03630 55	.77
.24	.05350 40	.82931 20	.26188 80	.03769 60	.76
.25	.05468 75	.82031 25	.27343 75	.03906 25	.75
.26	.05579 60	.81118 80	.28501 20	.04040 40	.74
.27	.05683 05	.80194 15	.29660 85	.04171 95	.73
.28	.05779 20	.79257 60	.30822 40	.04300 80	.72
.29	.05868 15	.78309 45	.31985 55	.04426 85	.71
.30	.05950 00	.77350 00	.33150 00	.04550 00	.70

TABLE XV B

FOUR-POINT INTERPOLATION COEFFICIENTS

p ($p<.5$)	c_{-1} −	c_0 +	c_1 +	c_2 +	p ($p>.5$)
.30	.05950 00	.77350 00	.33150 00	.04550 00	.70
.31	.06024 85	.76379 55	.34315 45	.04670 15	.69
.32	.06092 80	.75398 40	.35481 60	.04787 20	.68
.33	.06153 95	.74406 85	.36648 15	.04901 05	.67
.34	.06208 40	.73405 20	.37814 80	.05011 60	.66
.35	.06256 25	.72393 75	.38981 25	.05118 75	.65
.36	.06297 60	.71372 80	.40147 20	.05222 40	.64
.37	.06332 55	.70342 65	.41312 35	.05322 45	.63
.38	.06361 20	.69303 60	.42476 40	.05418 80	.62
.39	.06383 65	.68255 95	.43639 05	.05511 35	.61
.40	.06400 00	.67200 00	.44800 00	.05600 00	.60
.41	.06410 35	.66136 05	.45958 95	.05684 65	.59
.42	.06414 80	.65064 40	.47115 60	.05765 20	.58
.43	.06413 45	.63985 35	.48269 65	.05841 55	.57
.44	.06406 40	.62899 20	.49420 80	.05913 60	.56
.45	.06393 75	.61806 25	.50568 75	.05981 25	.55
.46	.06375 60	.60706 80	.51713 20	.06044 40	.54
.47	.06352 05	.59601 15	.52853 85	.06102 95	.53
.48	.06323 20	.58489 60	.53990 40	.06156 80	.52
.49	.06289 15	.57372 45	.55122 55	.06205 85	.51
.50	.06250 00	.56250 00	.56250 00	.06250 00	.50

SECTION 3. NORMAL PROBABILITY FUNCTIONS

The arguments p and q of Table XV C are the areas of the larger and the smaller tails of a unit normal distribution dichotomized at the point x, which is a standard score deviate. I is the area from the mean to this point x, and z is the ordinate at the point x.

The E^{II} footings at the bottoms of the columns of this table give the maximum linear interpolation errors for I, p, or q arguments half way down the columns. On the last two pages of this table E^{III}, the maximum three point interpolation error when I, p, or q is the argument, is given. The E^{-II} footings are the maximum inverse linear interpolation error in I, p, or q when x is the argument.

TABLE XV C

I	x	z	q	z/q	z/p	pq	p
.000	.000000	.398942	**.500**	.79788	.79788	.250000	**.500**
.001	.002507	.398941	.499	.79948	.79629	.249999	.501
.002	.005013	.398937	.498	.80108	.79470	.249996	.502
.003	.007520	.398931	.497	.80268	.79310	.249991	.503
.004	.010027	.398922	.496	.80428	.79151	.249984	.504
.005	.012533	.398911	.495	.80588	.78992	.249975	.505
.006	.015040	.398897	.494	.80748	.78833	.249964	.506
.007	.017547	.398881	.493	.80909	.78675	.249951	.507
.008	.020054	.398862	.492	.81070	.78516	.249936	.508
.009	.022562	.398841	.491	.81230	.78358	.249919	.509
.010	.025069	.398816	**.490**	.81391	.78199	.249900	**.510**
.011	.027576	.398791	.489	.81552	.78041	.249879	.511
.012	.030084	.398762	.488	.81714	.77883	.249856	.512
.013	.032592	.398730	.487	.81875	.77725	.249831	.513
.014	.035100	.398697	.486	.82036	.77567	.249804	.514
.015	.037608	.398660	.485	.82198	.77410	.249775	.515
.016	.040117	.398621	.484	.82360	.77252	.249744	.516
.017	.042626	.398580	.483	.82522	.77095	.249711	.517
.018	.045135	.398536	.482	.82684	.76938	.249676	.518
.019	.047644	.398490	.481	.82846	.76780	.249639	.519
.020	.050154	.398441	**.480**	.83008	.76623	.249600	**.520**
.021	.052664	.398389	.479	.83171	.76466	.249559	.521
.022	.055174	.398336	.478	.83334	.76309	.249516	.522
.023	.057684	.398279	.477	.83497	.76153	.249471	.523
.024	.060195	.398220	.476	.83660	.75996	.249424	.524
.025	.062707	.398159	.475	.83823	.75840	.249375	.525
.026	.065219	.398096	.474	.83986	.75683	.249324	.526
.027	.067731	.398028	.473	.84150	.75527	.249271	.527
.028	.070243	.397959	.472	.84292	.75371	.249216	.528
.029	.072756	.397888	.471	.84477	.75215	.249159	.529
.030	.075270	.397814	**.470**	.84641	.75059	.249100	**.530**
.031	.077784	.397737	.469	.84805	.74903	.249039	.531
.032	.080298	.397658	.468	.84970	.74748	.248976	.532
.033	.082813	.397577	.467	.85134	.74592	.248911	.533
.034	.085329	.397493	.466	.85299	.74437	.248844	.534
.035	.087845	.397406	.465	.85464	.74281	.248775	.535
.036	.090361	.397317	.464	.85629	.74126	.248704	.536
.037	.092878	.397225	.463	.85794	.73971	.248631	.537
.038	.095395	.397131	.462	.85959	.73816	.248556	.538
.039	.097914	.397034	.461	.86125	.73661	.248479	.539
.040	.100434	.396935	**.460**	.86290	.73506	.248400	**.540**
E^{11} .0000000+	.000000+ E^{-11}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.040	.100434	.396935	.460	.86290	.73506	.248400	.540
.041	.102953	.396834	.459	.86456	.73352	.248319	.541
.042	.105474	.396729	.458	.86622	.73197	.248236	.542
.043	.107995	.396623	.457	.86788	.73043	.248151	.543
.044	.110516	.396513	.456	.86955	.72888	.248064	.544
.045	.113039	.396401	.455	.87121	.72734	.247975	.545
.046	.115562	.396287	.454	.87288	.72580	.247884	.546
.047	.118085	.396170	.453	.87455	.72426	.247791	.547
.048	.120610	.396051	.452	.87622	.72272	.247696	.548
.049	.123135	.395929	.451	.87789	.72118	.247599	.549
.050	.125661	.395805	.450	.87957	.71964	.247500	.550
.051	.128188	.395678	.449	.88124	.71811	.247399	.551
.052	.130716	.395549	.448	.88292	.71657	.247296	.552
.053	.133245	.395417	.447	.88460	.71504	.247191	.553
.054	.135774	.395282	.446	.88628	.71350	.247084	.554
.055	.138304	.395145	.445	.88797	.71197	.246975	.555
.056	.140835	.395005	.444	.88965	.71044	.246864	.556
.057	.143367	.394863	.443	.89134	.70891	.246751	.557
.058	.145900	.394719	.442	.89303	.70738	.246636	.558
.059	.148434	.394572	.441	.89472	.70585	.246519	.559
.060	.150969	.394422	.440	.89641	.70432	.246400	.560
.061	.153505	.394270	.439	.89811	.70280	.246279	.561
.062	.156042	.394115	.438	.89981	.70127	.246156	.562
.063	.158580	.393957	.437	.90150	.69975	.246031	.563
.064	.161119	.393798	.436	.90321	.69822	.245904	.564
.065	.163658	.393635	.435	.90491	.69670	.245775	.565
.066	.166199	.393470	.434	.90661	.69518	.245644	.566
.067	.168741	.393303	.433	.90832	.69366	.245511	.567
.068	.171285	.393133	.432	.91003	.69214	.245376	.568
.069	.173829	.392960	.431	.91174	.69062	.245239	.569
.070	.176374	.392785	.430	.91345	.68910	.245100	.570
.071	.178921	.392608	.429	.91517	.68758	.244959	.571
.072	.181468	.392427	.428	.91689	.68606	.244816	.572
.073	.184017	.392245	.427	.91861	.68455	.244671	.573
.074	.186567	.392059	.426	.92033	.68303	.244524	.574
.075	.189118	.391870	.425	.92205	.68151	.244375	.575
.076	.191671	.391681	.424	.92378	.68000	.244224	.576
.077	.194225	.391488	.423	.92550	.67849	.244071	.577
.078	.196780	.391293	.422	.92723	.67698	.243916	.578
.079	.199336	.391095	.421	.92897	.67547	.243759	.579
.080	.201893	.390894	.420	.93070	.67396	.243600	.580
E^{11}	.000000+	000000+		.00000+	.00000+	.000000+	
.0000000+	E^{-11}						

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.080	.201893	.390894	**.420**	.93070	.67396	.243600	**.580**
.081	.204452	.390691	.419	.93244	.67245	.243439	.581
.082	.207013	.390485	.418	.93417	.67094	.243276	.582
.083	.209574	.390277	.417	.93592	.66943	.243111	.583
.084	.212137	.390066	.416	.93766	.66792	.242944	.584
.085	.214702	.389852	.415	.93940	.66641	.242775	.585
.086	.217267	.389636	.414	.94115	.66491	.242604	.586
.087	.219835	.389418	.413	.94290	:66340	.242431	.587
.088	.222403	.389197	.412	.94465	.66190	.242256	.588
.089	.224973	.388973	.411	.94641	66040	.242079	.589
.090	.227545	.388747	**.410**	.94816	.65889	.241900	**.590**
.091	.230118	.388518	.409	.94992	.65739	.241719	.591
.092	.232693	.388287	.408	.95168	.65589	.241536	.592
.093	.235269	.388053	.407	.95345	.65439	.241351	.593
.094	.237847	.387816	.406	.95521	.65289	.241164	.594
.095	.240426	.387577	.405	.95698	.65139	.240975	.595
.096	.243007	.387335	.404	.95875	.64989	.240784	.596
.097	.245590	.387091	.403	96052	.64839	.240591	.597
.098	.248174	.386844	.402	.96230	.64690	.240396	.598
.099	.250760	.386595	.401	.96408	.64540	.240199	.599
.100	253347	.386342	**.400**	.96586	.64390	.240000	**.600**
.101	.255936	.386088	.399	.96764	.64241	.239799	.601
.102	.258527	.385831	.398	.96942	.64091	.239596	.602
.103	.261120	.385571	.397	.97121	.63942	.239391	.603
.104	.263714	.385308	.396	.97300	.63793	.239184	.604
.105	.266311	.385043	.395	.97479	.63749	.238975	.605
.106	.268909	.384776	.394	.97659	.63494	.238764	.606
.107	.271508	.384506	.393	.97839	.63345	.238551	.607
.108	.274110	.384233	.392	.98019	.63196	.238336	.608
.109	.276714	.383957	.391	.98199	.63047	.238119	.609
.110	.279319	.383679	**.390**	.98379	.62898	.237900	**.610**
.111	.281926	.383399	.389	.98560	.62749	.237679	.611
.112	.284536	.383115	.388	.98741	.62600	.237456	.612
.113	.287147	.382830	.387	.98922	.62452	.237231	.613
.114	.289760	.382541	.386	.99104	.62303	.237004	.614
.115	.292375	.382250	.385	.99286	.62154	.236775	.615
.116	.294992	.381956	.384	.99468	.62006	.236544	.616
.117	.297611	.381660	.383	.99650	.61857	.236311	.617
.118	.300232	.381361	.382	.99833	.61709	.236076	.618
.119	.302855	.381060	.381	1.00016	.61561	.235839	.619
.120	.305481	.380755	**.380**	1.00199	.61412	.235600	**.620**
E^{11}	.000000+	.000000+		.00000+	.00000+	.000000+	
.0000001	E^{-11}						

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.120	.305481	.380755	.380	1.00199	.61412	.235600	.620
.121	.308108	.380449	.379	1.00382	.61264	.235359	.621
.122	.310738	.380139	.378	1.00566	.61116	.235116	.622
.123	.313369	.379827	.377	1.00750	.60967	.234871	.623
.124	.316003	.379513	.376	1.00934	.60819	.234624	.624
.125	.318639	.379195	.375	1.01119	.60671	.234375	.625
.126	.321278	.378875	.374	1.01303	.60523	.234124	.626
.127	.323918	.378553	.373	1.01489	.60375	.233871	.627
.128	.326561	.378227	.372	1.01674	.60227	.233616	.628
.129	.329206	.377900	.371	1.01860	.60079	.233359	.629
.130	.331853	.377569	.370	1.02046	.59932	.233100	.630
.131	.334503	.377236	.369	1.02232	.59784	.232839	.631
.132	.337155	.376900	.368	1.02418	.59636	.232576	.632
.133	.339809	.376562	.367	1.02605	.59488	.232311	.633
.134	.342466	.376220	.366	1.02792	.59341	.232044	.634
.135	.345125	.375877	.365	1.02980	.59193	.231775	.635
.136	.347787	.375530	.364	1.03168	.59046	.231504	.636
.137	.350451	.375181	.363	1.03356	.58898	.231231	.637
.138	.353118	.374829	.362	1.03544	.58751	.230956	.638
.139	.355787	.374475	.361	1.03733	.58603	.230679	.639
.140	.358459	.374118	.360	1.03922	.58456	.230400	.640
.141	.361133	.373758	.359	1.04111	.58309	.230119	.641
.142	.363810	.373395	.358	1.04300	.58161	.229836	.642
.143	.366489	.373030	.357	1.04490	.58014	.229551	.643
.144	.369171	.372662	.356	1.04680	.57867	.229264	.644
.145	.371856	.372292	.355	1.04871	.57720	.228975	.645
.146	.374544	.371919	.354	1.05062	.57573	.228684	.646
.147	.377234	.371543	.353	1.05253	.57426	.228391	.647
.148	.379927	.371164	.352	1.05444	.57278	.228096	.648
.149	.382622	.370783	.351	1.05636	.57131	.227799	.649
.150	.385320	.370399	.350	1.05828	.56984	.227500	.650
.151	.388022	.370012	.349	1.06021	.56837	.227199	.651
.152	.390726	.369623	.348	1.06214	.56691	.226896	.652
.153	.393433	.369231	.347	1.06407	.56544	.226591	.653
.154	.396142	.368836	.346	1.06600	.56397	.226284	.654
.155	.398855	.368439	.345	1.06794	.56250	.225975	.655
.156	.401571	.368038	.344	1.06988	.56103	.225664	.656
.157	.404289	.367635	.343	1.07182	.55957	.225351	.657
.158	.407011	.367230	.342	1.07377	.55810	.225036	.658
.159	.409735	.366821	.341	1.07572	.55663	.224719	.659
.160	.412463	.366410	.340	1.07768	.55517	.224400	.660
E^{11} .000001	.000000+ E^{-11}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.160	.412463	.366410	**.340**	1.07768	.55517	.224400	**.660**
.161	.415194	.365996	.339	1.07963	.55370	.224079	.661
.162	.417928	.365580	.338	1.08160	.55224	.223756	.662
.163	.420665	.365160	.337	1.08356	.55077	.223431	.663
.164	.423405	.364738	.336	1.08553	.54930	.223104	.664
.165	.426148	.364314	.335	1.08750	.54784	.222775	.665
.166	.428895	.363886	.334	1.08948	.54638	.222444	.666
.167	.431644	.363456	.333	1.09146	.54491	.222111	.667
.168	.434397	.363023	.332	1.09344	.54345	.221776	.668
.169	.437154	.362587	.331	1.09543	.54198	.221439	.669
.170	.439913	.362149	**.330**	1.09742	.54052	.221100	**.670**
.171	.442676	.361707	.329	1.09941	.53906	.220759	.671
.172	.445443	.361263	.328	1.10141	.53759	.220416	.672
.173	.448212	.360817	.327	1.10342	.53613	.220071	.673
.174	.450986	.360367	.326	1.10542	.53467	.219724	.674
.175	.453762	.359915	.325	1.10743	.53321	.219375	.675
.176	.456542	.359459	.324	1.10944	.53174	.219024	.676
.177	.459326	.359001	.323	1.11146	.53028	.218671	.677
.178	.462113	.358541	.322	1.11348	.52882	.218316	.678
.179	.464904	.358077	.321	1.11550	.52736	.217959	.679
.180	.467699	.357611	**.320**	1.11753	.52590	.217600	**.680**
.181	.470497	.357142	.319	1.11957	.52444	.217239	.681
.182	.473299	.356670	.318	1.12160	.52298	.216876	.682
.183	.476104	.356195	.317	1.12364	.52152	.216511	.683
.184	.478914	.355718	.316	1.12569	.52006	.216144	.684
.185	.481727	.355237	.315	1.12774	.51859	.215775	.685
.186	.484544	.354754	.314	1.12979	.51713	.215404	.686
.187	.487365	.354268	.313	1.13185	.51567	.215031	.687
.188	.490189	.353780	.312	1.13391	.51422	.214656	.688
.189	.493018	.353288	.311	1.13597	.51275	.214279	.689
.190	.495850	.352793	**.310**	1.13804	.51129	.213900	**.690**
.191	.498687	.352296	.309	1.14012	.50984	.213519	.691
.192	.501527	.351796	.308	1.14219	.50838	.213136	.692
.193	.504372	.351293	.307	1.14428	.50692	.212751	.693
.194	.507221	.350787	.306	1.14636	.50546	.212364	.694
.195	.510073	.350279	.305	1.14846	.50400	.211975	.695
.196	.512930	.349767	.304	1.15055	.50254	.211584	.696
.197	.515792	.349253	.303	1.15265	.50108	.211191	.697
.198	.518657	.348736	.302	1.15475	.49962	.210796	.698
.199	.521527	.348216	.301	1.15686	.49816	.210399	.699
.200	.524401	.347693	**.300**	1.15898	.49670	.210000	**.700**
E^{11} .0000002	.000000+ E^{-11}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.200	.524401	.347693	.300	1.15898	.49670	.210000	.700
.201	.527279	.347167	.299	1.16109	.49525	.209599	.701
.202	.530161	.346638	.298	1.16321	.49379	.209196	.702
.203	.533048	.346107	.297	1.16534	.49233	.208791	.703
.204	.535940	.345572	.296	1.16747	.49087	.208384	.704
.205	.538836	.345035	.295	1.16961	.48941	.207975	.705
.206	.541737	.344494	.294	1.17175	.48795	.207564	.706
.207	.544642	.343951	.293	1.17389	.48649	.207151	.707
.208	.547551	.343405	.292	1.17604	.48504	.206736	.708
.209	.550466	.342856	.291	1.17820	.48358	.206319	.709
.210	.553385	.342304	.290	1.18036	.48212	.205900	.710
.211	.556308	.341749	.289	1.18252	.48066	.205479	.711
.212	.559237	.341191	.288	1.18469	.47920	.205056	.712
.213	.562170	.340631	.287	1.18687	.47774	.204631	.713
.214	.565108	.340067	.286	1.18904	.47628	.204204	.714
.215	.568051	.339500	.285	1.19123	.47483	.203775	.715
.216	.570999	.338931	.284	1.19342	.47337	.203344	.716
.217	.573952	.338358	.283	1.19561	.47191	.202911	.717
.218	.576910	.337783	.282	1.19781	.47045	.202476	.718
.219	.579873	.337205	.281	1.20002	.46899	.202039	.719
.220	.582841	.336623	.280	1.20223	.46753	.201600	.720
.221	.585815	.336039	.279	1.20444	.46607	.201159	.721
.222	.588793	.335452	.278	1.20666	.46461	.200716	.722
.223	.591777	.334861	.277	1.20888	.46315	.200271	.723
.224	.594766	.334268	.276	1.21112	.46170	.199824	.724
.225	.597760	.333672	.275	1.21335	.46024	.199375	.725
.226	.600760	.333073	.274	1.21559	.45878	.198924	.726
.227	.603765	.332470	.273	1.21784	.45732	.198471	.727
.228	.606775	.331865	.272	1.22009	.45586	.198016	.728
.229	.609791	.331257	.271	1.22235	.45440	.197559	.729
.230	.612813	.330646	.270	1.22461	.45294	.197100	.730
.231	.615840	.330031	.269	1.22688	.45148	.196639	.731
.232	.618873	.329414	.268	1.22916	.45002	.196176	.732
.233	.621912	.328793	.267	1.23143	.44856	.195711	.733
.234	.624956	.328170	.266	1.23372	.44710	.195244	.734
.235	.628006	.327544	.265	1.23601	.44564	.194775	.735
.236	.631062	.326914	.264	1.23831	.44418	.194304	.736
.237	.634124	.326281	.263	1.24061	.44272	.193831	.737
.238	.637192	.325646	.262	1.24292	.44125	.193356	.738
.239	.640265	.325007	.261	1.24524	.43979	.192879	.739
.240	.643345	.324365	.260	1.24756	.43833	.192400	.740
E^{11} .0000002	.000001 E^{-11}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.240	.643345	.324365	**.260**	1.24756	.43833	.192400	**.740**
.241	.646431	.323720	.259	1.24988	.43687	.191919	.741
.242	.649524	.323072	.258	1.25222	.43541	.191436	.742
.243	.652622	.322421	.257	1.25456	.43394	.190951	.743
.244	.655727	.321767	.256	1.25690	.43248	.190464	.744
.245	.658838	.321110	.255	1.25925	.43102	.189975	.745
.246	.661955	.320449	.254	1.26161	.42956	.189484	.746
.247	.665079	.319786	.253	1.26398	.42809	.188991	.747
.248	.668209	.319119	.252	1.26635	.42663	.188496	.748
.249	.671346	.318449	.251	1.26872	.42517	.187999	.749
.250	.674490	.317777	**.250**	1.27111	.42370	.187500	**.750**
.251	.677640	.317101	.249	1.27350	.42224	.186999	.751
.252	.680797	.316421	.248	1.27589	.42077	.186496	.752
.253	.683961	.315739	.247	1.27830	.41931	.185991	.753
.254	.687131	.315053	.246	1.28070	.41784	.185484	.754
.255	.690309	.314365	.245	1.28312	.41638	.184975	.755
.256	.693493	.313673	.244	1.28554	.41491	.184464	.756
.257	.696685	.312978	.243	1.28798	.41345	.183951	.757
.258	.699884	.312279	.242	1.29041	.41198	.183436	.758
.259	.703090	.311578	.241	1.29285	.41051	.182919	.759
.260	.706303	.310873	**.240**	1.29531	.40904	.182400	**.760**
.261	.709523	.310165	.239	1.29776	.40758	.181879	.761
.262	.712751	.309454	.238	1.30023	.40611	.181356	.762
.263	.715986	.308740	.237	1.30270	.40464	.180831	.763
.264	.719229	.308022	.236	1.30518	.40317	.180304	.764
.265	.722479	.307301	.235	1.30767	.40170	.179775	.765
.266	.725737	.306577	.234	1.31016	.40023	.179244	.766
.267	.729003	.305850	.233	1.31266	.39876	.178711	.767
.268	.732276	.305119	.232	1.31517	.39729	.178176	.768
.269	.735558	.304385	.231	1.31768	.39582	.177639	.769
.270	.738847	.303648	**.230**	1.32021	.39435	.177100	**.770**
.271	.742144	.302908	.229	1.32274	.39288	.176559	.771
.272	.745450	.302164	.228	1.32528	.39140	.176016	.772
.273	.748763	.301417	.227	1.32783	.38993	.175471	.773
.274	.752085	.300666	.226	1.33038	.38846	.174924	.774
.275	.755415	.299913	.225	1.33294	.38698	.174375	.775
.276	.758754	.299155	.224	1.33551	.38551	.173824	.776
.277	.762101	.298395	.223	1.33809	.38403	.173271	.777
.278	.765456	.297631	.222	1.34068	.38256	.172716	.778
.279	.768820	.296864	.221	1.34328	.38108	.172159	.779
.280	.772193	.296094	**.220**	1.34588	.38069	.171600	**.780**
E^{11} .0000003	.000001 E^{-11}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.280	.772193	.296094	.220	1.34588	.38069	.171600	.780
.281	.775575	.295320	.219	1.34849	.37813	.171039	.781
.282	.778966	.294542	.218	1.35111	.37665	.170476	.782
.283	.782365	.293762	.217	1.35374	.37517	.169911	.783
.284	.785774	.292978	.216	1.35638	.37370	.169344	.784
.285	.789192	.292190	.215	1.35902	.37222	.168775	.785
.286	.792619	.291399	.214	1.36168	.37074	.168204	.786
.287	.796055	.290605	.213	1.36434	.36926	.167631	.787
.288	.799501	.289807	.212	1.36701	.36778	.167056	.788
.289	.802956	.289006	.211	1.36970	.36629	.166479	.789
.290	.806421	.288201	.210	1.37239	.36481	.165900	.790
.291	.809896	.287393	.209	1.37509	.36333	.165319	.791
.292	.813380	.286582	.208	1.37780	.36185	.164736	.792
.293	.816875	.285766	.207	1.38051	.36036	.164151	.793
.294	.820379	.284948	.206	1.38324	.35888	.163564	.794
.295	.823894	.284126	.205	1.38598	.35739	.162975	.795
.296	.827418	.283300	.204	1.38873	.35590	.162384	.796
.297	.830953	.282471	.203	1.39148	.35442	.161791	.797
.298	.834499	.281638	.202	1.39425	.35293	.161196	.798
.299	.838055	.280802	.201	1.39702	.35144	.160599	.799
.300	.841621	.279962	.200	1.39981	.34995	.160000	.800
.301	.845199	.279118	.199	1.40260	.34846	.159399	.801
.302	.848787	.278272	.198	1.40541	.34697	.158796	.802
.303	.852386	.277421	.197	1.40823	.34548	.158191	.803
.304	.855996	.276567	.196	1.41106	.34399	.157584	.804
.305	.859617	.275709	.195	1.41389	.34250	.156975	.805
.306	.863250	.274847	.194	1.41674	.34100	.156364	.806
.307	.866894	.273982	.193	1.41960	.33951	.155751	.807
.308	.870550	.273114	.192	1.42247	.33801	.155136	.808
.309	.874217	.272241	.191	1.42535	.33652	.154519	.809
.310	.877896	.271365	.190	1.42824	.33502	.153900	.810
.311	.881587	.270486	.189	1.43114	.33352	.153279	.811
.312	.885291	.269602	.188	1.43405	.33202	.152656	.812
.313	.889006	.268715	.187	1.43698	.33052	.152031	.813
.314	.892733	.267824	.186	1.43991	.32902	.151404	.814
.315	.896473	.266929	.185	1.44286	.32752	.150775	.815
.316	.900226	.266031	.184	1.44582	.32602	.150144	.816
.317	.903991	.265129	.183	1.44879	.32452	.149511	.817
.318	.907770	.264223	.182	1.45177	.32301	.148876	.818
.319	.911561	.263313	.181	1.45477	.32151	.148239	.819
.320	.915365	.262400	.180	1.45778	.32000	.147600	.820
E^{ii} .0000004	.000001 E^{-ii}	.000000+		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.320	.915365	.262400	.180	1.45778	.32000	.147600	.820
.321	.919183	.261483	.179	1.46080	.31849	.146959	.821
.322	.923014	.260562	.178	1.46383	.31699	.146316	.822
.323	.926859	.259637	.177	1.46688	.31548	.145671	.823
.324	.930717	.258708	.176	1.46993	.31397	.145024	.824
.325	.934589	.257775	.175	1.47300	.31245	.144375	.825
.326	.938476	.256839	.174	1.47609	.31094	.143724	.826
.327	.942376	.255898	.173	1.47918	.30943	.143071	.827
.328	.946291	.254954	.172	1.48229	.30792	.142416	.828
.329	.950221	.254006	.171	1.48541	.30640	.141759	.829
.330	.954165	.253054	.170	1.48855	.30488	.141100	.830
.331	.958125	.252097	.169	1.49170	.30337	.140439	.831
.332	.962099	.251137	.168	1.49486	.30185	.139776	.832
.333	.966088	.250173	.167	1.49804	.30033	.139111	.833
.334	.970093	.249205	.166	1.50123	.29881	.138444	.834
.335	.974114	.248233	.165	1.50444	.29728	.137775	.835
.336	.978150	.247257	.164	1.50766	.29576	.137104	.836
.337	.982203	.246277	.163	1.51090	.29424	.136431	.837
.338	.986271	.245292	.162	1.51415	.29271	.135756	.838
.339	.990356	.244304	.161	1.51742	.29118	.135079	.839
.340	.994458	.243312	.160	1.52070	.28966	.134400	.840
.341	.998576	.242315	.159	1.52399	.28813	.133719	.841
.342	1.002712	.241315	.158	1.52731	.28660	.133036	.842
.343	1.006864	.240310	.157	1.53064	.28507	.132351	.843
.344	1.011034	.239301	.156	1.53398	.28353	.131664	.844
.345	1.015222	.238288	.155	1.53734	.28200	.130975	.845
.346	1.019428	.237270	.154	1.54071	.28046	.130284	.846
.347	1.023651	.236249	.153	1.54411	.27892	.129591	.847
.348	1.027893	.235223	.152	1.54752	.27739	.128896	.848
.349	1.032154	.234193	.151	1.55095	.27585	.128199	.849
.350	1.036433	.233159	.150	1.55439	.27430	.127500	.850
.351	1.040732	.232120	.149	1.55785	.27276	.126799	.851
.352	1.045050	.231077	.148	1.56133	.27122	.126096	.852
.353	1.049387	.230030	.147	1.56483	.26967	.125391	853
.354	1.053744	.228979	.146	1.56835	.26813	.124684	.854
.355	1.058122	.227923	.145	1.57188	.26658	.123975	.855
.356	1.062519	.226862	.144	1.57543	.26503	.123264	.856
.357	1.066938	.225798	.143	1.57901	.26347	.122551	.857
.358	1.071377	.224728	.142	1.58259	.26192	.121836	.858
.359	1.075837	.223655	.141	1.58621	.26037	.121119	.859
.360	1.080319	.222577	.140	1.58983	.25881	.120400	.860
E^{11} .0000006	.000002 E^{-11}	.000001		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.360	1.080319	.222577	.140	1.58983	.25881	.120400	.860
.361	1.084823	.221494	.139	1.59348	.25725	.119679	.861
.362	1.089349	.220407	.138	1.59715	.25569	.118956	.862
.363	1.093897	.219315	.137	1.60084	.25413	.118231	.863
.364	1.098468	.218219	.136	1.60455	.25257	.117504	.864
.365	1.103063	.217119	.135	1.60829	.25100	.116775	.865
.366	1.107680	.216013	.134	1.61204	.24944	.116044	.866
.367	1.112321	.214903	.133	1.61581	.24787	.115311	.867
.368	1.116987	.213789	.132	1.61961	.24630	.114576	.868
.369	1.121676	.212669	.131	1.62343	.24473	.113839	.869
.370	1.126391	.211545	.130	1.62727	.24316	.113100	.870
.371	1.131131	.210416	.129	1.63113	.24158	.112359	.871
.372	1.135896	.209283	.128	1.63502	.24000	.111616	.872
.373	1.140688	.208145	.127	1.63894	.23842	.110871	.873
.374	1.145505	.207001	.126	1.64286	.23684	.110124	.874
.375	1.150349	.205853	.125	1.64683	.23526	.109375	.875
.376	1.155221	.204701	.124	1.65081	.23368	.108624	.876
.377	1.160120	.203543	.123	1.65482	.23209	.107871	.877
.378	1.165047	.202380	.122	1.65885	.23050	.107116	.878
.379	1.170002	.201213	.121	1.66292	.22891	.106359	.879
.380	1.174987	.200040	.120	1.66700	.22732	.105600	.880
.381	1.180001	.198863	.119	1.67112	.22572	.104839	.881
.382	1.185044	.197680	.118	1.67525	.22413	.104076	.882
.383	1.190118	.196493	.117	1.67943	.22253	.103311	.883
.384	1.195223	.195300	.116	1.68362	.22093	.102544	.884
.385	1.200359	.194102	.115	1.68785	.21932	.101775	.885
.386	1.205527	.192900	.114	1.69211	.21772	.101004	.886
.387	1.210727	.191691	.113	1.69638	.21611	.100231	.887
.388	1.215960	.190478	.112	1.70070	.21450	.099456	.888
.389	1.221227	.189259	.111	1.70504	.21289	.098679	.889
.390	1.226528	.188036	.110	1.70941	.21128	.097900	.890
.391	1.231864	.186806	.109	1.71382	.20966	.097119	.891
.392	1.237235	.185572	.108	1.71826	.20804	.096336	.892
.393	1.242642	.184332	.107	1.72273	.20642	.095551	.893
.394	1.248085	.183087	.106	1.72724	.20480	.094764	.894
.395	1.253565	.181836	.105	1.73177	.20317	.093975	.895
.396	1.259084	.180579	.104	1.73634	.20154	.093184	.896
.397	1.264641	.179318	.103	1.74095	.19991	.092391	.897
.398	1.270237	.178050	.102	1.74559	.19827	.091596	.898
.399	1.275874	.176777	.101	1.75027	.19664	.090799	.899
.400	1.281552	.175498	.100	1.75498	.19500	.090000	.900
E^{ii} .0000007	.000003 E^{-ii}	.000001		.00000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.400	1.281552	.175498	.100	1.7550	.19500	.090000	.900
.401	1.287271	.174214	.099	1.7597	.19336	.089199	.901
.402	1.293032	.172924	.098	1.7645	.19171	.088396	.902
.403	1.298837	.171628	.097	1.7694	.19006	.087591	.903
.404	1.304685	.170326	.096	1.7742	.18841	.086784	.904
.405	1.310579	.169018	.095	1.7791	.18676	.085975	.905
.406	1.316519	.167705	.094	1.7841	.18510	.085164	.906
.407	1.322505	.166385	.093	1.7891	.18345	.084351	.907
.408	1.328539	.165060	.092	1.7941	.18178	.083536	.908
.409	1.334622	.163728	.091	1.7992	.18012	.082719	.909
.410	1.340755	.162391	.090	1.8043	.17845	.081900	.910
.411	1.346939	.161047	.089	1.8095	.17678	.081079	.911
.412	1.353174	.159697	.088	1.8147	.17511	.080256	.912
.413	1.359463	.158340	.087	1.8200	.17343	.079431	.913
.414	1.365806	.156978	.086	1.8253	.17175	.078604	.914
.415	1.372204	.155609	.085	1.8307	.17006	.077775	.915
.416	1.378659	.154233	.084	1.8361	.16838	.076944	.916
.417	1.385172	.152851	.083	1.8416	.16669	.076111	.917
.418	1.391744	.151463	.082	1.8471	.16499	.075276	.918
.419	1.398377	.150068	.081	1.8527	.16329	.074439	.919
.420	1.405072	.148666	.080	1.8582	.16159	.073600	.920
.421	1.411830	.147258	.079	1.8640	.15989	.072759	.921
.422	1.418654	.145843	.078	1.8698	.15818	.071916	.922
.423	1.425544	.144420	.077	1.8756	.15647	.071071	.923
.424	1.432503	.142991	.076	1.8815	.15475	.070224	.924
.425	1.439531	.141555	.075	1.8874	.15303	.069375	.925
.426	1.446632	.140112	.074	1.8934	.15131	.068524	.926
.427	1.453806	.138662	.073	1.8995	.14958	.067671	.927
.428	1.461056	.137205	.072	1.9056	.14785	.066816	.928
.429	1.468384	.135740	.071	1.9118	.14611	.065959	.929
.430	1.475791	.134268	.070	1.9181	.14437	.065100	.930
.431	1.483280	.132788	.069	1.9245	.14263	.064239	.931
.432	1.490853	.131301	.068	1.9309	.14088	.063376	.932
.433	1.498513	.129807	.067	1.9374	.13913	.062511	.933
.434	1.506262	.128304	.066	1.9440	.13737	.061644	.934
.435	1.514102	.126794	.065	1.9507	.13561	.060775	.935
.436	1.522036	.125276	.064	1.9574	.13384	.059904	.936
.437	1.530068	.123750	.063	1.9643	.13207	.059031	.937
.438	1.538199	.122216	.062	1.9712	.13029	.058156	.938
.439	1.546433	.120674	.061	1.9783	.12851	.057279	.939
.440	1.554774	.119123	.060	1.9854	.12673	.056400	.940
E^{ii} .000001	.000008 E^{-ii}	.000001		.0000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.440	1.554774	.119123	.060	1.9854	.12673	.056400	.940
.441	1.563224	.117564	.059	1.9926	.12494	.055519	.941
.442	1.571787	.115996	.058	1.9999	.12314	.054636	.942
.443	1.580467	.114420	.057	2.0074	.12134	.053751	.943
.444	1.589268	.112836	.056	2.0149	.11953	.052864	.944
.445	1.598193	.111242	.055	2.0226	.11772	.051975	.945
.446	1.607248	.109639	.054	2.0304	.11590	.051084	.946
.447	1.616436	.108027	.053	2.0382	.11407	.050191	.947
.448	1.625763	.106406	.052	2.0463	.11224	.049296	.948
.449	1.635234	.104776	.051	2.0544	.11041	.048399	.949
.450	1.644854	.103136	.050	2.0627	.10856	.047500	.950
.451	1.654628	.101486	.049	2.0711	.10672	.046599	.951
.452	1.664563	.099826	.048	2.0797	.10486	.045696	.952
.453	1.674665	.098157	.047	2.0884	.10300	.044791	.953
.454	1.684941	.096477	.046	2.0973	.10113	.043884	.954
.455	1.695398	.094787	.045	2.1064	.09925	.042975	.955
.456	1.706044	.093086	.044	2.1156	.09737	.042064	.956
.457	1.716886	.091375	.043	2.1250	.09548	.041151	.957
.458	1.727934	.089652	.042	2.1346	.09358	.040236	.958
.459	1.739198	.087919	.041	2.1444	.09168	.039319	.959
.460	1.750686	.086174	.040	2.1544	.08976	.038400	.960
.461	1.762410	.084417	.039	2.1645	.08784	.037479	.961
.462	1.774382	.082649	.038	2.1750	.08591	.036556	.962
.463	1.786614	.080868	.037	2.1856	.08398	.035631	.963
.464	1.799118	.079075	.036	2.1965	.08203	.034704	.964
.465	1.811911	.077270	.035	2.2077	.08007	.033775	.965
.466	1.825007	.075452	.034	2.2192	.07811	.032844	.966
.467	1.838424	.073620	.033	2.2309	.07613	.031911	.967
.468	1.852180	.071775	.032	2.2430	.07415	.030976	.968
.469	1.866296	.069915	.031	2.2553	.07215	.030039	.969
.470	1.880794	.068042	.030	2.2681	.07015	.029100	.970
.471	1.895698	.066154	.029	2.2812	.06813	.028159	.971
.472	1.911036	.064250	.028	2.2946	.06610	.027216	.972
.473	1.926837	.062332	.027	2.3086	.06406	.026271	.973
.474	1.943134	.060397	.026	2.3230	.06201	.025324	.974
.475	1.959964	.058445	.025	2.3378	.05994	.024375	.975
.476	1.977368	.056476	.024	2.3532	.05786	.023424	.976
.477	1.995393	.054490	.023	2.3691	.05577	.022471	.977
.478	2.014091	.052485	.022	2.3857	.05367	.021516	.978
.479	2.033520	.050462	.021	2.4030	.05154	.020559	.979
.480	2.053749	.048418	.020	2.4209	.04941	.019600	.980
E^{ii} E^{iii} .000003	.00003 .000000+ E^{-ii}	.000001		.0000+	.00000+	.000000+	

TABLE XV C (CONTINUED)

I	x	z	q	z/q	z/p	pq	p
.480	2.053749	.048418	**.020**	2.4209	.04941	.019600	**.980**
.481	2.074855	.046354	.019	2.4397	.04725	.018639	.981
.482	2.096927	.044268	.018	2.4593	.04508	.017676	.982
.483	2.120072	.042160	.017	2.4800	.04289	.016711	.983
.484	2.144411	.040028	.016	2.5018	.04068	.015744	.984
.485	2.170090	.037870	.015	2.5247	.03845	.014775	.985
.486	2.197286	.035687	.014	2.5491	.03619	.013804	.986
.487	2.226211	.033475	.013	2.5750	.03392	.012831	.987
.488	2.257129	.031234	.012	2.6028	.03161	.011856	.988
.489	2.290370	.028960	.011	2.6327	.02928	.010879	.989
.490	2.326348	.026652	**.010**	2.665	.02692	.009900	**.990**
.491	2.365618	.024306	.009	2.701	.02453	.008919	.991
.492	2.408916	.021920	.008	2.740	.02210	.007936	.992
.493	2.457264	.019487	.007	2.784	.01962	.006951	.993
.494	2.512144	.017003	.006	2.834	.01711	.005964	.994
.495	2.575829	.014460	.005	2.892	.01453	.004975	.995
.496	2.652070	.011847	.004	2.962	.01189	.003984	.996
.497	2.747781	.009149	.003	3.050	.00918	.002991	.997
.498	2.878161	.006340	.002	3.170	.00635	.001996	.998
.499	3.090229	.003367	.001	3.367	.00337	.000999	.999
E^{ii}	.0005	.000005		.000+	.00000+	.000000+	
E^{iii}	.00005						
.00001	E^{-ii}						

SECTION 4. SQUARE AND CUBE ROOTS

The methods given in Chapter XIII, Section 13, for extracting square and cube roots on a computing machine are particularly expeditious when a good first approximation is started with. The two-figure argument table herewith provides such, especially if the approximate root is gotten by linear interpolation.

TABLE XV D

SQUARE AND CUBE ROOTS

N	$\sqrt{N}$	$\sqrt{10N}$	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
1.0	1.00000	3.16228	1.00000	2.15443	4.64159
1.1	1.04881	3.31662	1.03228	2.22398	4.79142
1.2	1.09545	3.46410	1.06266	2.28943	4.93242
1.3	1.14018	3.60555	1.09139	2.35133	5.06580
1.4	1.18322	3.74166	1.11869	2.41014	5.19249
1.5	1.22474	3.87298	1.14471	2.46621	5.31329
1.6	1.26491	4.00000	1.16961	2.51984	5.42884
1.7	1.30384	4.12311	1.19348	2.57128	5.53966
1.8	1.34164	4.24264	1.21644	2.62074	5.64622
1.9	1.37840	4.35890	1.23856	2.66840	5.74890
2.0	1.41421	4.47214	1.25992	2.71442	5.84804
2.1	1.44914	4.58258	1.28058	2.75892	5.94392
2.2	1.48324	4.69042	1.30059	2.80204	6.03681
2.3	1.51658	4.79583	1.32001	2.84387	6.12693
2.4	1.54919	4.89898	1.33887	2.88450	6.21447
2.5	1.58114	5.00000	1.35721	2.92402	6.29961
2.6	1.61245	5.09902	1.37507	2.96250	6.38250
2.7	1.64317	5.19615	1.39248	3.00000	6.46330
2.8	1.67332	5.29150	1.40946	3.03659	6.54213
2.9	1.70294	5.38516	1.42604	3.07232	6.61911
3.0	1.73205	5.47723	1.44225	3.10723	6.69433

MAXIMUM INTERPOLATION ERRORS IN THE NEIGHBORHOOD OF THE INDICATED N'S

		$\sqrt{N}$	$\sqrt{10N}$	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
1.0	E^{ii}	.00031	.00099	.00028	.00060	.00130
	E^{iii}	.00002	.00007	.00002	.00004	.00020
2.0	E^{ii}	.00011	.00035	.00009	.00019	.00041
	E^{iii}		.00001		.00001	.00002
3.0	E^{ii}	.00007	.00020	.00005	.00010	.00022
	E^{iii}					.00001

TABLE XV D (Continued)

SQUARE AND CUBE ROOTS

N	$\sqrt{N}$	$\sqrt{10N}$	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
3.0	1.73205	5.47723	1.44225	3.10723	6.69433
3.1	1.76068	5.56776	1.45810	3.14138	6.76790
3.2	1.78885	5.65685	1.47361	3.17480	6.83990
3.3	1.81659	5.74456	1.48881	3.20753	6.91042
3.4	1.84391	5.83095	1.50369	3.23961	6.97953
3.5	1.87083	5.91608	1.51829	3.27107	7.04730
3.6	1.89737	6.00000	1.53262	3.30193	7.11379
3.7	1.92354	6.08276	1.54668	3.33222	7.17905
3.8	1.94936	6.16441	1.56049	3.36198	7.24316
3.9	1.97484	6.24500	1.57406	3.39121	7.30614
4.0	2.00000	6.32456	1.58740	3.41995	7.36806
4.1	2.02485	6.40312	1.60052	3.44822	7.42896
4.2	2.04939	6.48074	1.61343	3.47603	7.48887
4.3	2.07364	6.55744	1.62613	3.50340	7.54784
4.4	2.09762	6.63325	1.63864	3.53035	7.60590
4.5	2.12132	6.70820	1.65096	3.55689	7.66309
4.6	2.14476	6.78233	1.66310	3.58305	7.71944
4.7	2.16795	6.85565	1.67507	3.60883	7.77498
4.8	2.19089	6.92820	1.68687	3.63424	7.82974
4.9	2.21359	7.00000	1.69850	3.65931	7.88374
5.0	2.23607	7.07107	1.70998	3.68403	7.93701

MAXIMUM INTERPOLATION ERRORS IN THE NEIGHBORHOOD OF THE INDICATED N'S

3.0, E''	.00007	.00020	.00005	.00010	.00022
3.0, E'''					.00001
4.0, E''	.00004	.00008	.00003	.00006	.00013
5.0, E''	.00003	.00009	.00002	.00004	.00009

TABLE XV D (Continued)

SQUARE AND CUBE ROOTS

N	$\sqrt{N}$	$\sqrt{10N}$	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
5.0	2.23607	7.07107	1.70998	3.68403	7.93701
5.1	2.25832	7.14143	1.72130	3.70843	7.98957
5.2	2.28035	7.21110	1.73248	3.73251	8.04145
5.3	2.30217	7.28011	1.74351	3.75629	8.09267
5.4	2.32379	7.34847	1.75441	3.77976	8.14325
5.5	2.34521	7.41620	1.76517	3.80295	8.19321
5.6	2.36643	7.48331	1.77581	3.82586	8.24257
5.7	2.38747	7.54983	1.78632	3.84850	8.29134
5.8	2.40832	7.61577	1.79670	3.87088	8.33955
5.9	2.42899	7.68115	1.80697	3.89300	8.38721
6.0	2.44949	7.74597	1.81712	3.91487	8.43433
6.1	2.46982	7.81025	1.82716	3.93650	8.48093
6.2	2.48998	7.87401	1.83709	3.95789	8.52702
6.3	2.50998	7.93725	1.84691	3.97906	8.57262
6.4	2.52982	8.00000	1.85664	4.00000	8.61774
6.5	2.54951	8.06226	1.86626	4.02073	8.66239
6.6	2.56905	8.12404	1.87578	4.04124	8.70659
6.7	2.58844	8.18535	1.88520	4.06155	8.75034
6.8	2.60768	8.24621	1.89454	4.08166	8.79366
6.9	2.62679	8.30662	1.90378	4.10157	8.83656
7.0	2.64575	8.36660	1.91293	4.12129	8.87904

MAXIMUM INTERPOLATION ERRORS IN THE NEIGHBORHOOD OF THE INDICATED N'S

5.0, E^{ii}	.00003	.00009	.00002	.00004	.00009
6.0, E^{ii}	.00002	.00007	.00001	.00003	.00007
7.0, E^{ii}	.00002	.00005	.00001	.00002	.00005

TABLE XV D (Continued)

SQUARE AND CUBE ROOTS

N	$\sqrt{N}$	$\sqrt{10N}$	$\sqrt[3]{N}$	$\sqrt[3]{10N}$	$\sqrt[3]{100N}$
7.0	2.64575	8.36660	1.91293	4.12129	8.87904
7.1	2.66458	8.42615	1.92200	4.14082	8.92112
7.2	2.68328	8.48528	1.93098	4.16017	8.96281
7.3	2.70185	8.54400	1.93988	4.17934	9.00411
7.4	2.72029	8.60233	1.94870	4.19834	9.04504
7.5	2.73861	8.66025	1.95743	4.21716	9.08560
7.6	2.75681	8.71780	1.96610	4.23582	9.12581
7.7	2.77489	8.77496	1.97468	4.25432	9.16566
7.8	2.79285	8.83176	1.98319	4.27266	9.20516
7.9	2.81069	8.88819	1.99163	4.29084	9.24434
8.0	2.82843	8.94427	2.00000	4.30887	9.28318
8.1	2.84605	9.00000	2.00830	4.32675	9.32170
8.2	2.86356	9.05539	2.01653	4.34448	9.35990
8.3	2.88097	9.11043	2.02469	4.36207	9.39780
8.4	2.89828	9.16515	2.03279	4.37952	9.43539
8.5	2.91548	9.21954	2.04083	4.39683	9.47268
8.6	2.93258	9.27362	2.04880	4.41400	9.50969
8.7	2.94958	9.32738	2.05671	4.43105	9.54640
8.8	2.96648	9.38083	2.06456	4.44796	9.58284
8.9	2.98329	9.43398	2.07235	4.46475	9.61900
9.0	3.00000	9.48683	2.08008	4.48140	9.65489
9.1	3.01662	9.53939	2.08776	4.49794	9.69052
9.2	3.03315	9.59166	2.09538	4.51436	9.72589
9.3	3.04959	9.64365	2.10294	4.53065	9.76100
9.4	3.06594	9.69536	2.11045	4.54684	9.79586
9.5	3.08221	9.74679	2.11791	4.56290	9.83048
9.6	3.09839	9.79796	2.12532	4.57886	9.86485
9.7	3.11448	9.84886	2.13267	4.59470	9.89898
9.8	3.13050	9.89949	2.13997	4.61044	9.93288
9.9	3.14643	9.94987	2.14723	4.62607	9.96655
10.0	3.16228	10.00000	2.15443	4.64159	10.00000
7.0, E^{ii}	.00002	.00005	.00001	.00002	.00005
8.0, E^{ii}	.00001	.00004	.00001	.00002	.00004
9.0, E^{ii}	.00001	.00004	.00001	.00001	.00003
10.0 E^{ii}	.00001	.00003	.00001	.00001	.00003

SECTION 5. SHORTER MATHEMATICAL TABLES LISTED ACCORDING TO OCCURRENCE IN EARLIER CHAPTERS

APPENDIX A

MATHEMATICAL BACKGROUND TEST

SECTION 1. THE BACKGROUND TEST

Herein is given a "Background Test for Elementary Statistics." The purpose of the test is threefold, *(a)* to enable a student to secure an idea as to whether his mathematical background is adequate for the pursuit of an elementary course in statistics, *(b)* to inform him of shortcomings in his background, if they exist, which he should remedy by study of the appropriate elementary topics as found in arithmetics and elementary algebras, and *(c)* to enable an instructor to classify his students and properly adapt instruction to them. The two first mentioned purposes may be realized by the student himself without the aid or knowledge of the instructor. To secure information *(a)* it is necessary that the student take the test in an entirely fair manner. He must observe time limits which, however, are not short. He must not coach himself upon the items of the test by consulting others who know the content of the test, or by reference to the scoring key given on pages following the test. He must be rigorous in marking his own paper. It may happen that the student's answer will be different from that given in the key, which is the sole guide for marking papers,

but different in a manner which the student realizes, though none other would, is not due to a faulty understanding. In this case, he should mark his answer wrong in spite of his personal knowledge that he correctly understands the problem and its solution. If he is not thus objective and rigorous in the grading of his own paper, he cannot interpret his position by comparing his score with the percentile scores given.

Page 1 of the test, calling for certain personal information about the student, provides a second somewhat different and somewhat less reliable means of estimating a student's preparation for pursuing elementary statistics. Items *a, c, d, e,* and *f* are probably pertinent but are not involved in the experience score as calculated from this page. Their inclusion on the page will help the student to a rounded out picture of himself and of his equipment. The items which are scored and combined into an experience score are X_1, X_2, X_3 as explained herewith.

X_1 is a score based on item *(b)*, year in college, according to the following schedule:

X_1 score	Year in college
1	Freshman
2	Sophomore
3	Junior
4	Senior
5	First year of graduate study
6	Second year of graduate study
7	Third year of graduate study

X_2 is a score based on item *(g)*, the most advanced course in mathematics successfully pursued, according to the following schedule:

X_1 score	Most advanced course in mathematics
8	Arithmetic, i.e., no algebra
9	1/2 year or 1 year of High School algebra

10	Other high school mathematics, including a second course in algebra, trigonometry or geometry.
11	College algebra, or answer omitted
12	Calculus
13	One course beyond calculus
14	Two or more courses beyond calculus

X_2 is a score based on related courses and quality of work according to the following schedule:

$$\text{Item (h) credit} \begin{cases} -1 & \text{for mark of D or F} \\ 0 & \text{for mark of C or I or no answer} \\ 1 & \text{for mark of A or B} \end{cases}$$

$$\text{Item (i) credit} \begin{cases} 0 & \text{for no course} \\ 1 & \text{for one or more courses (or un-} \\ & \text{doubted equivalent in research work)} \end{cases}$$

$$\text{Item (j) credit} \begin{cases} -1 & \text{for mark of D or F} \\ 0 & \text{for mark of C or I or no answer} \\ 1 & \text{for mark of A or B} \end{cases}$$

X_2 = sum of credits

The final experience score X_e is given by

$$X_e = 2X_1 + 2X_2 + X_2$$

A BACKGROUND TEST FOR ELEMENTARY STATISTICS

Parts I, II

FILL IN FULLY THE INFORMATION ASKED FOR ON THIS PAGE

(a) Name ______________ Institution __________ Date __________

(b) If an undergraduate student, encircle the year in college. 1 2 3 4 If a graduate student, encircle the year of graduate work. 1 2 3

(c) Candidate for a degree? Yes___No___If so, for what degree?______________________________

(d) Degree (or degrees) held__________ Institution __________ ______________________________ Date __________

(e) Teaching, research, or administrative position now held, if any? ______________________________

____

(f) Experimental, measurement, and statistical duties connected therewith, if any? ______________

Courses and experience connected with measurement, and statistics:

(g) Most advanced course in mathematics pursued____________ Date______________ Institution ______________

(h) Using the marking system: A = highest quarter; B = next to highest quarter; C = next to lowest quarter; D = lowest quarter; F = failure; I = incomplete, course not finished; give yourself a mark to indicate as nearly as you can your standing in the course just mehtioned

(i) Course or courses in measurement and statistics________ Date______________ Institution ______________

(j) As above, for each course, in the order mentioned, give yourself a mark A, B, C, D, F, or I____________

SUMMARY

		Equivalent percentile (Beginning students)
Experience		
Parts I and II (equivalent score)		
Parts III, IV, and V (equivalent score)		
Total: Score on Pts. I, II, III, IV, and V		

The purpose of this test is to enable an instructor of statistics to estimate the sort of work which the students of his class may pursue profitably, and to enable students of statistics in general to appraise their own mathematical equipment. Some students may be prepared for advanced work. Therefore, the test contains items which many students will be unable to do who are promising candidates for an elementary course. No student should feel discouraged, and each one should do all of those items which lie within his capacity. Work as rapidly as is consistent with accuracy.

INSTRUCTIONS TO ADMINISTRATOR OF TEST:

After 35 minutes working time announce, "Even if you have not finished Part I start on Part II."

After another 7 minutes, or a total working time of 42 minutes, call time.

An intermission is recommended before starting Parts III, IV, and V.

SUMMARY

PART	NO. RIGHT	WEIGHT, OR MULTIPLYING FACTOR	(NO. RT.) X (WT.)
I		1	
II		1/6	
	CRUDE TOTAL		
PARTS I AND II SCORE FROM TABLE BELOW			

Crude Total	0	1	2	3	4	5	6	7	8	9	10	11
Equivalent Score	3	4	5	6	8	9	10	11	12	13	14	15
Crude Total	12	13	14	15	16	17	18	19	20	21	22	23
Equivalent Score	16	17	19	20	21	22	23	24	25	26	27	28

PART I. COMPUTATION

A. Addition and Subtraction

Carry all answers to the greatest number of decimal places found in the data of the problem, unless otherwise specified. Designate the answers clearly.

1. Add 32.784, 764.46, 1.2733, .0056.

2. Add 73.8, −13.4, .632, −.076.

3. Subtract 17.384 from 8.2631.

4. Subtract .0758 from 1.6.

5. Subtract −.848 from 3 3/4.

B. Multiplication

Carry answers to as many decimal places as the sum of the decimal places in the two factors. When the common fractions given are changed to decimals express them to the two nearest decimal places.

6. Multiply 73.18 by .062.

7. Multiply 13 1/6 by 4.74.

8. Multiply −1.14 by 14.32.

9. Multiply −.037 by 7/16.

10. Multiply 8/14 by 10/9. (Express answer as a decimal to 3 places)

C. Division

Carry answers to three significant figures.

11. Divide .1145 by 3.76.

12. Divide 9 7/12 by 24.28.

13. Divide 1 3/4 by 72 4/9.

D. Roots and Powers

Carry square root answers to as many significant figures as given in the number the square root of which is to be extracted.

14. Extract the square root of 5776.

15. Extract the square root of .00049.

16. Extract the square root of −169.

17. Expand $(9)^2$.

E. Percentage

Keep answers to the nearest integer.

18. Find 52 percent of 165.

19. Find .52 percent of 916.

20. The ratio of 3 to 4 equals the ratio of 7.5 to what?

(Go right on to Part II)

PART II. EXPONENTS

Simplify the following:

1. $(4)^3\,(4)^2 =$ ____________

2. $x^4x^2 =$ ____________

3. $x^5(-x^3) =$ ____________

4. $2(x^4y^2)^2 =$ ____________

5. $\dfrac{x^5}{x^3} =$ ____________

6. $\dfrac{x^7\;\;y^6}{x^4\;\;y^2} =$ ____________

7. $x^y\;\;x^2 =$ ____________

8. $m^a\;\;m =$ ____________

9. $\dfrac{m^5\;\;n^4}{m^7\;\;n^7} =$ ____________

10. $(4x^3)^2 =$ ____________

11. $(x^5)^2 =$ ____________

12. $\sqrt{y^6} =$ ____________

13. $\sqrt{x^{2c}} =$ ____________

14. $\sqrt[4]{x^{12}y^2} =$ ____________

15. $(x^{a-1})^2 =$ ____________

16. $\sqrt[c]{Z^{3c}} =$ ____________

17. $m^{-4}m^5 =$ ____________

18. $\dfrac{n^6}{n^{-4}} =$ ____________

19. $x^{-4}x^{-2} =$ ____________

20. $\dfrac{3x^3(c-d)^3}{-x^2(c-d)^2}$ ____________

(Go right on to Part III)

A BACKGROUND TEST FOR ELEMENTARY STATISTICS

PARTS III, IV, V

Name____________________ Institution ____________________

Date ____________________

INSTRUCTIONS TO ADMINISTRATOR OF TEST:

After 35 minutes working time announce, "Even if you have not finished Part III start on Part IV."

After another 2 minutes announce, "Even if you have not finished Part IV start on Part V."

After another 15 minutes, or a total working time of 52 minutes, collect all booklets.

PART	NO. RIGHT	WT.	(NO. RT.) X (WT.)
III		1	
IV		1/3	
V, Ex. 1-6		1/6	
V, Ex. 7		1/3	
V, Ex. 8		2/3	
V, Ex. 9-10		1	
V, Ex. 11		2	
V, Ex. 12		4	
TOTAL PART V			
CRUDE TOTAL PARTS III, IV, AND V			
PARTS III, IV, AND V SCORE FROM TABLE BELOW			

CRUDE TOTAL	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
EQUIVALENT SCORE	8	9	10	11	11	12	12	13	13	14	14	15	16	16	17	17	18	18
CRUDE TOTAL	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	
EQUIVALENT SCORE	19	19	20	21	21	22	22	23	23	24	24	25	26	26	27	27	28	

PART III. ALGEBRA

1. Add $5x + 8$ and $3x + 4y - 3$.

2. Add $4x^2 - 3x + xy - 4$ and $3x^2 + 6x - y^2 - 2x^2 - 2$.

3. Subtract $4x^4 - 5x^2 + 3x - 5$ from $x^5 + 9x^4 - 2x + 3$.

4. Multiply $4x^2$ by $7x$.

5. Multiply $4x^2 + 3$ by $-4xy$.

6. Multiply $4x^2 - 3x - 5$ by $2x - 4$.

7. Expand $(m - n)^2$.

8. Factor $m^2 - n^2$.

9. Write from memory the first 4 terms of $(m + n)^a$. (Do not derive)

10. Solve the equation $6x - 12 = 0$.

11. What are the roots of $x^2 - 6x + 8 = 0$?

12. What are the roots of $6x^2 + 2x - 20 = 0$?

13. The number of real or imaginary roots of $5x^3 + 3x^2 + x - 4 = 0$ is ____________________

14. Given the formula $t = \frac{a^2}{4b}$. What value will t have if $a = 8$ and $b = 2$?

15. $w = 1 - \frac{5\Sigma P^2}{K(K^2 - 3)}$. If $(\Sigma P^2) = 30$, and $K = 5$, solve for w to two decimal places.

16. $D = \frac{A}{G_1 + \frac{G_2}{H}}$. If $D = 25$, $A = 170$, $G_1 = 4$, $G_2 = 84$, solve for H.

17. The two sides of a right triangle are 8 and 1.8. What is the length of the hypotenuse?

18. The quotient of two numbers is y and divisor is n. Represent the dividend.

19. $ax^2 + bx + c = 0$ is a typical quadratic equation, the solution of which is given by

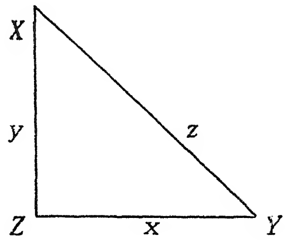
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Express the roots of $5x^2 + 2x - 4 = 0$ by means of this formula, but do not perform the indicated arithmetic computation.

20. Solve simultaneously
$$2x - 3y = 7$$
$$3x + y = 5$$

(Go right on to Part IV)

PART IV. TRIGONOMETRY



1. What is the sine of angle X in the adjacent figure? ________________

2. The cosine of angle Y? __________

3. The tangent of angle Y? __________

4. What is the numerical value of the tangent of 45 degrees? __________

____

5. The cosine of 0 degrees is ______

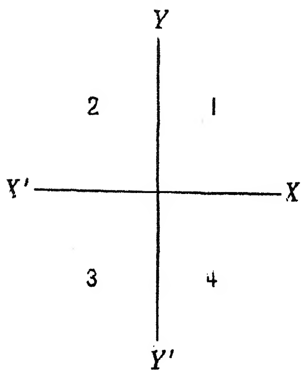
6. What is the greatest possible value of the sine of an angle? ________

____

(Go right on to Part V)

PART V. ANALYTIC GEOMETRY

1-6. If the quadrants made by perpendicular coordinate axes are numbered as in the diagram, indicate as result of inspection, in the column headed quadrant, the number of the quadrant in which each of the following points lies.



x	y	quadrant
−2	5	
−4	−6	
2	3	
9	−3	
5	−6	
−2	−7	

7. A straight line is represented by an equation in x and y of the ______________ degree.

8. The line $5x + 4y = 12$ intercepts the Y axis at what point? ______________________________

9. The slope of the line $y = 5x - 3$ is ______________

10. The slope of the line $5x - 3y = 4$ is ______________

11. The slope of the line passing through the point $(x = 6,\ y = 7)$ and the point $(x = 2,\ y = 5)$ is ______________________________

12. $4Y + 5X = -10$ is the equation of a line. Shifting axes 2 units to the right and 2 units up, and using x and y to designate the new variables, what then is the equation of the line? ______________________________

SECTION 2. SCORING KEYS

A BACKGROUND TEST FOR ELEMENTARY STATISTICS

PARTS I, II

PART I: COMPUTATION
EACH ITEM WEIGHTED 1

1. 798.5229
2. 60.956
3. −9.1209
4. 1.5242
5. 4.598
6. 4.53716
7. 62.4258 or 62.41
8. −16.3248
9. −.01628 or −.0162 or −.016
10. .635 or .634 or .633
11. .030 or .03 or .0304 or .0305
12. .395 or .394
13. .024 or .0241
14. 76. or 76 or 76.00 or 76.0
15. .022 or .02214
16. $13\sqrt{-1}$ or $13i$
17. 81
18. 86. ($\frac{1}{2}$ credit for 85.80)
19. 5. ($\frac{1}{2}$ credit for 4.7632)
20. 10

PART II: EXPONENTS
EACH ITEM WEIGHTED 1/6

1. 4^5 or 1024
2. x^6
3. $-x^8$
4. $2x^8y^4$
5. x^2
6. x^3y^4
7. $x^{(y+2)}$
8. m^{a+1}
9. $m^{-2}n^{-3}$ or $\frac{1}{m^2\,n^3}$
10. $16\,x^6$ or $(4)^2\,x^6$
11. x^{10}
12. y^3 or $\pm\,y^3$
13. x^c or $\pm\,x^c$
14. $x^2y^{\frac{1}{2}}$ or $x^2\sqrt{y}$ or $\pm$ either
15. x^{2a-2} or $x^{2(a-1)}$
16. z^3
17. m
18. n^{10}
19. x^{-6} or $\frac{1}{x^6}$
20. $-3x\,(c-d)$ or $-3xc+3xd$

SCORING KEY

A BACKGROUND TEST FOR ELEMENTARY STATISTICS

PART III

PART III: ALGEBRA
EACH ITEM WEIGHTED 1

1. $8x + 4y + 5$
2. $5x^2 + 3x + xy - y^2 - 6$
3. $x^5 + 5x^4 + 5x^2 - 5x + 8$
4. $28x^3$
5. $-16x^3y - 12xy$
6. $8x^3 - 22x^2 + 2x + 20$
7. $m^2 - mn + n^2$
8. $(m - n)(m + n)$
9. $m^a + am^{(a-1)}n + \dfrac{a(a-1)}{2(\text{or } \lfloor 2 \text{ or } 2!)} m^{(a-2)}n^2 + \dfrac{a(a-1)(a-2)}{6(\text{or } \lfloor 3 \text{ or } 3!)} m^{(a-3)}n^3 \ldots\ldots$
10. 2
11. 2, 4
12. $\frac{5}{3}$, (or 1 2/3), -2
13. 3
14. 8
15. $-.36$, or $\frac{-4}{11}$
16. 30
17. 8.2 or $\sqrt{67.24}$
18. ny
19. $\dfrac{-2 \pm \sqrt{84}}{10}$ or $\dfrac{-2 \pm \sqrt{4 + 80}}{10}$
20. 2, -1

SCORING KEY

A BACKGROUND TEST FOR ELEMENTARY STATISTICS

PARTS IV, V

PART IV: TRIGONOMETRY

EACH ITEM WEIGHTED 1/3

1. $\frac{x}{z}$

2. $\frac{x}{z}$

3. $\frac{y}{x}$

4. 1

5. 1

6. 1

PART V: ANALYTIC GEOMETRY

1-6. 2
3
1 (each weighted 1/6)
4
4
3

7. 1st or linear (weight 1/3)

8. 3 or $y = 3$ (weight 2/3)

9. 5 (weight 1)

10. $\frac{5}{3}$ (weight 1)

11. .5 or $\frac{1}{2}$ (weight 2)

12. $4y + 5x = -28$ (weight 4)

SECTION 3. PERCENTILE NORMS AND RELIABILITY

Percentile equivalents are given in Table Appendix A for the Experience Score and also for the Total Background Test Score, based on 406 students beginning work in statistics from Columbia University Teachers College, University of Illinois, Harvard University, University of Minnesota, University of Oregon, and Stanford University. Percentiles for the Parts separately are not provided, but the "equivalent" scores are such that a person's Part I and II score will equal his Part III, IV, and V score if he is strictly typical or equally able in the subject matter of these different parts. In this table, X_b is the Background Test Score and X_e the Experience Score. The percentiles are based upon 406 students beginning elementary statistics in courses at *(a)* Columbia University (Teachers College, $N = 52$), *(b)* University of Illinois ($N = 39$), *(c)* Harvard University (Graduate School of Education ($N = 29$), *(d)* University of Minnesota ($N = 88$), *(e)* University of Oregon ($N = 153$), *(f)* Stanford University ($N = 45$).

The standard error of X_e as a measure of whatever it is a measure is unknown because a second similar measure is not available nor can the elements entering into X_e be divided into two comparable halves and reliability determined by the Spearman-Brown formula. The standard error of X_b when taken as evidence of $\tilde{X}_b$ is 1.9. This is equivalent to a reliability coefficient of .80 in the group of 406, having a standard deviation of 4.40. The correlation between X_e and X_b for the 406 cases is .65. Certain small samples suggest that the correlations between a term grade in an elementary course in statistics and X_e and X_b scores, received four months earlier, are in the neighborhood of .4 and .6 respectively.

TABLE Appx. A. PERCENTILES

PERCENTILES	X_b BACKGROUND TEST SCORE	X_e EXPERIENCE SCORE	PERCENTILES	X_b BACKGROUND TEST SCORE	X_e EXPERIENCE SCORE
$P_{.01}$	11	22	$P_{.55}$	27	32
$P_{.02}$	12	23	$P_{.60}$	28	32
$P_{.05}$	14	24	$P_{.65}$	30	33
$P_{.10}$	16	24	$P_{.70}$	31	34
$P_{.15}$	17	25	$P_{.75}$	32	34
$P_{.20}$	18	26	$P_{.80}$	34	35
$P_{.25}$	19	26	$P_{.85}$	37	36
$P_{.30}$	21	27	$P_{.90}$	39	37
$P_{.35}$	22	28	$P_{.95}$	42	38
$P_{.40}$	24	29	$P_{.98}$	45	40
$P_{.45}$	24	30	$P_{.99}$	46	41
$P_{.50}$	25	31			

APPENDIX B

REFERENCE LISTS

SECTION 1. LIST OF COMMON STATISTICAL SYMBOLS

Herein are given the more common meanings attached to symbols as used in this text, and also the more common meanings as used in statistical texts in general.

Throughout, as used in this text, a bar over a symbol, e.g., $\bar{V}$, indicates an unbiased estimate or a mean value, and a tilde, e.g., $\tilde{V}$, a population, hypothetical, or true value.

THE GREEK ALPHABET

A	α	Alpha	I	ι	Iota	P	ρ	Rho
B	β	Beta	K	κ	Kappa	Σ	σ	Sigma
Γ	γ	Gamma	Λ	λ	Lambda	T	τ	Tau
Δ	δ	Delta	M	μ	Mu	Υ	υ	Upsilon
E	ϵ	Epsilon	N	ν	Nu	Φ	ϕ	Phi
Z	ζ	Zeta	Ξ	ξ	Xi	X	χ	Chi
H	η	Eta	O	o	Omicron	Ψ	ψ	Psi
Θ	θ	Theta	Π	π	Pi	Ω	ω	Omega

LIST OF COMMON STATISTICAL SYMBOLS — LITERAL SYMBOLS

a, A, and α

a,- the constant term in a regression equation.

a,- a and b are employed in connection with boundary values of a sequential analysis likelihood ratio.

a_{ij},- an element in a matrix at the intersection of the i-th row and the j-th column, Ch. XIV, Sec. 2. [14:01]

A,- Arbitrary Origin, as employed by Yule and Kendall.

A_{ij},- a cofactor. Ch. XIV, Sec. 2. [14:12]

$\mathcal{A}$,- common notation for a matrix. Ch. XIV, Sec. 2. [14:01]

$\mathcal{A}'$,- a matrix that is the transpose of $\mathcal{A}$. [14:02]

$\mathcal{A}^{-1}$,- a matrix that is the inverse of $\mathcal{A}$. [14:06]

α,- the proportion of cases in a cell in a 2X2-fold. The other proportions are β, γ, and δ.

α,- in Pearson's *Tables*, the area between mean and point of dichotomy in a unit normal distribution $= \alpha/2$.

α,- a measure of risk in sequential analysis.

b, B, and β

b,- a regression coefficient.

b,- see a (sequential analysis).

b_i,- an abridged notation for $b_{0i.12\ldots)i(\ldots k}$. A regression coefficient in a multiple regression equation.

β,- a standard score regression coefficient.

β,- a measure of risk in sequential analysis.

β_i,- an abridged notation for $b_{0i.12\ldots)i(\ldots k}$. A regression coefficient in a multiple standard score regression equation.

$B(p,q)$,- The Beta Function. See Table XV A. 1934-1938 *Biom.*

$B_x(p,q)$,- The incomplete Beta Function. See Table XV A. 1934-1938 *Biom.*
See $I_x(p,q)$.

$\beta_1 = \mu_3^2/\mu_2^3$; $\beta_2 = \mu_4/\mu_2^2$; $\beta_3 = \mu_3\mu_5/\mu_2^4$; $\beta_4 = \mu_6/\mu_2^3$

$\beta_5 = \mu_3\mu_7/\mu_2^5$; $\beta_6 = \mu_8/\mu_2^4$, in which the μ's are moments from the mean, and $\mu_2 = V$ as used in this text.

These β's are used especially in connection with the Pearson system of curves.

c, C (for γ and Γ see "g"; for χ see end of alphabet)

c,- as a preceding subscript indicates a correction for coarseness of grouping.

c_{ij},- the covariance, or product moment, between the i and j variables.

c_{-3}, c_{-2}, c_{-1}, c_0, c_1, c_2, c_3, c_4,- Lagrangian interpolation coefficients.

C,- a contingency coefficient: $C^2 = \phi^2/(1+\phi^2)$
$= \chi^2/(N + \chi^2)$

C_{ij},- a cofactor.

C_{ij},- is used in this text to mean $\Sigma X_i X_j/N$.
See c_{ij}.

cm,- as a preceding subscript indicates a correction for use of class means.

$C(n,m)$, or ${}_nC_m$, or C^n_m, or $\binom{n}{m}$,- is the number of combinations of n things m at a time.
$C(n,m) = m!/[m!(n-m)!]$. See $P(n,m)$.

d, D, δ, and Δ

d,- a deviation.

d,- a difference in scores, in ranks, etc.

d,- (and also d_{ij}) the mean deviation of a portion of a unit normal distribution. [8:26] and [8:27]

d_{ij},- a difference between standard scores z_i and z_j. [11:28]

d_x,- in actuarial statistics is the number dying during the x year of life.

d.o.f.,- the number of degrees of freedom.

D,- a difference in rank when computing the correlation between ranks.

δ,- See first definition of α. δ is value tabled in Pearson's *Tables* for different tetrachoric coefficients.

δ and Δ frequently indicate small increments. As they approach infinitesimals they approach the differentials of calculus.

δ_{ij},- Kronecker's δ_{ij} is such a quantity that it = 1 when the two subscripts are the same and otherwise it = 0. If the a's in the matrix

$$\begin{Vmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1k} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2k} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ a_{k1} & a_{k2} & \cdot & \cdot & \cdot & a_{kk} \end{Vmatrix}$$

are such that the sum of their squares for any row, or column, = 1, and the sum of the products for corresponding members for any two columns, = 0, then

$$\delta_{ij} = \sum_{g=1}^{g=k} a_{ig} a_{jg}.$$

Δ as a subscript means "dependent upon." It is the opposite of the dot as a subscript, which means "independent of." It is employed herein in connection with scores, correlation coefficients, regression coefficients, variances, covariances, etc.

Δ,- a major determinant. Δ_{ij} is the minor determinant after row i and column j have been deleted from Δ. Ch. XIV, Sec. 2, [14:16].

Δ,- with superscripts is used to indicate differences in tabled values. Ch. XIII, Sec. 12. Table XIII I.

e, E, and ϵ

e,- the Naperian base of logarithms. It = 2.71828 18284 59045, etc.

exp,- in an equation, the expression following "exp" is the exponent of e.

E,- the expected value, the expectancy, or the mean value.

ϵ,- a small quantity.

ϵ,- an unbiased correlation ratio.

f, F, ϕ, and Φ

f,- a frequency, or number of cases.

f,- as a preceding subscript indicates a correction for fineness of grouping.

$f(x)$,- a function of x.

$f'(x)$,- the first derivative of $f(x)$,- thus, if $y = f(x)$, then $dy/dx = f'(x)$; f'',- the second derivative, etc.

$f_{ij} = \sqrt[3]{F_{ij}}$. Ch. IX, Sec. 5, [9:22].

f_{ij} = the frequency in the cell at the intersection of the i row and the j column.

F,- when computing percentiles, F = the sum of f's up to a certain point.

F_{ij},- a variance ratio having i d.o.f. in the numerator and j d.o.f. in the denominator. As herein used, the denominator variance is the error variance. An alternative and common usage is so to write F_{ij} that it > 1.

ϕ,- a common designation for an angle.

ϕ,- $\phi(x)$ a function of x.

ϕ,- Yule's ϕ is a product-moment correlation coefficient in a 2X2-fold.

ϕ^2,- is the mean square contingency. It $= \chi^2/N$. This ϕ^2 = Yule's ϕ squared in case of a 2X2-fold only.

g, G, γ, and Γ

g,- as used by Fisher, $g_1 = k_3/k_2^{3/2}$ and $g_2 = k_4/k_2^2$. These are sample estimates of his population parameter, γ_1 and γ_2.

$G(r,\nu)$ integral. See Ch. XV, Sec. 1. Pearson, (1914).

γ,- Fisher's $\gamma_1 = \mu_3/\mu_2^{3/2}$, and $\gamma_2 = (\mu_4/\mu_2^2)-3$. As used by Fisher, Greek letters indicate population, not sample values.

Γ,- the gamma function,- a concept of factorial extended to numbers other than integers. See $I_{x,p}$.

h, H, and η

h,- see k.

h,- h_1 and h_2 are intercepts in sequential analysis.

H,- hypothesis. H with following notation commonly means under the hypothesis indicated by the following notation.

η,- a correlation ratio. See ϵ.

i, I, and ι

i,- the imaginary quantity $\sqrt{-1}$. This universal use of i has not been called for in this text.

i,- the size of the grouping interval.

i or r,- an interest rate.

i and j as subscripts of V and in other connections, are used herein to indicate the numbers of d.o.f. Herein i' and j', as subscripts of V, indicate the variances of the i and j variables,—the prime being used to avoid conflict with the use of i and j to indicate d.o.f.

i as a subscript of a variable, or connected with a summation sign, stands for any of a number of variables, usually from 1 to k. j has a similar meaning for a second series of variables, usually from a to ℓ.

I,- the area from the mean to the point x in the unit normal distribution. I of Table XV C = $\alpha/2$ of Sheppard's table, given in Pearson's *Tables*.

I,- the identity matrix: $AA^{-1} = I$. Ch. XIV, Sec. 2, [14:05].

$I_{x,p}$ of Pearson's *Tables of the Incomplete Gamma Function* = $\Gamma_x(p+1) / \Gamma(p+1)$.

$I_x(p,q)$ of Pearson's *Tables of the Incomplete Beta Function* = $B_x(p,q) / B(p,q)$.

j and J

j,- see i.

$J_m(x)$,- a common designation of a Bessel function of the first kind.

k, K, and κ

k,- is a coefficient of alienation. It $= \sqrt{1-r^2}$.

k,- is the number of classes in a categorical series. Also the specific designation of the k-th class. ℓ is used with a similar meaning if a second series is involved.

k_1, k_2, k_3, k_4, as used by Fisher, are unbiased sample estimates of cumulants κ_1, κ_2, κ_3, and κ_4.

κ,- Fisher's cumulants are population statistics. $\kappa_1 = \tilde{M}$; $\kappa_2 = \tilde{V}$; $\kappa_3 = \tilde{\mu}_3$; $\kappa_4 + 3\kappa_2^2 = \tilde{\mu}_4$ in which the tilde measures are as used in this text.

ℓ, L, λ, and Λ

ℓ,- a direction cosine, usually in n-dimensional space.

ℓ,- see the second definition of k.

ℓ_x,- in actuarial statistics is the number of persons living at age x.

L,- a likelihood ratio.

log,- also log_{10},- logarithm to the base 10.

ln,- also log_e,- logarithm to the base e.

λ,- a direction cosine, usually in n-dimensional space.

λ,- a Lagrange multiplier. [14:118]

Λ,- in the expression "Λ-shaped" indicates a unimodal distribution.

m, M, and μ

M,- the arithmetic mean.

μ,- a moment from the mean: $\mu_1 = (\Sigma x/N) = 0$); $\mu_2 = (\Sigma x^2/N) = V = \sigma^2$; $\mu_3 = (\Sigma x^3/N)$; $\mu_4 = (\Sigma x^4/N)$; etc.

μ',- a moment from zero: $\mu_1' = \Sigma X/N$; $\mu_2' = \Sigma X^2/N$; $\mu_3' = \Sigma X^3/N$; $\mu_4' = \Sigma X^4/N$; etc.

n, N, and ν

n,- d.o.f., i.e., n is synonymous with i as used herein.

n,- a common designation of the last item in a series,- i.e., n is synonymous with k as used herein.

n,- a common designation of the exponent of a binomial, or polynomial.

n,- number of cases in a sample (though N is more frequently so used).

$\bar{n}$,- average sample number.

N,- the number of cases in a sample.

ν,- Greek nu, sometimes used with the meaning of μ' preceding.

o and O

o,- as a subscript in connection with index numbers, it commonly refers to the basal date.

o,- as a subscript in H_0 indicates the null hypothesis.

o,- a common designation of the criterion variable.

OC,- operating characteristic curve.

p, P, π, and Π

p,- a probability, i.e., a proportion of cases. If but two classes, q is the other proportion, so that $p + q = 1$. If k classes, $p_1 + p_2 + \ldots + p_k = 1$. Some writers employ $p + q + r + s = 1$, in lieu of $p_1 + p_2 + p_3 + p_4 = 1$.

p,- a price index.

p,- in interpolation, the proportionate distance of the value in question from the tabled argument next smaller to the distance between the two neighboring arguments.

p,- in a dichotomized unit normal distribution, p, as herein used, is the proportion to the left and q that to the right of the point of dichotomy. As tabled in this text, $p>q$.

p_{ij},- a product-moment, usually involving deviations from means. Frequently p_{ij} is defined as is c_{ij} herein. Another usage defines p_{ijkl} as $= \Sigma x_1^i x_2^j x_3^k x_4^l / N$, in which the exponents may take any positive integral values, including that of zero. Karl Pearson has used $\overline{p}_{ijkl}$ to mean p_{ijkl}, as here defined.

p_{ij},- a price ratio,—the price at date i divided by the price at date j. See q_{ij}.

p_x,- in actuarial statistics, the probability of a person of age x living one year.

P,- a probability, usually of a divergence as great in absolute value as that observed. With one d.o.f. and a normal distribution, $P = 2q$, as given in the first definition under p.

P,- a price index.

P_{ij},- a product moment. Sometimes P_{ij} is defined as C_{ij} herein. As herein used, $P_{ijkl} = \Sigma X_1^i X_2^j X_3^k X_4^l / N$. Karl Pearson uses p_{ijkl} to mean P_{ijkl}, as here defined.

$P(n,m)$, or ${}_nP_m$, or P_m^n,—is the number of permutations of n things m at a time. $P(n,m) = n!/(n-m)!$. See $C(n,m)$.

P_p,- a percentile, — P_p is the ($100p$) percentile.

π,- the ratio of circumference of circle to diameter. It = 3.14159 26535 89793 etc.

Π,- as a symbol of operation means the product of all those magnitudes immediately following, thus $\Pi_1^k X_i = X_1 \times X_2 \times \ldots \times X_k$.

q, Q

q,- see first and third definitions of p.

q_{ij},- a quantity ratio, — the quantity at date i divided by the quantity at date j. See p_{ij}.

q_{ij},- the area in a unit normal distribution between x_i and x_j.

q_x,- in actuarial statistics is the probability of a person of age x dying within one year. $q_x = d_x/\ell_x$.

Q,- a quantity index.

Q,- a quotient.

Q,- a Lexian ratio, which, when i = d.o.f., = χ^2/i = the variance ratio $F_{i,\infty}$.

r, R, and ρ

r,- a product-moment correlation coefficient.

r,- with preceding and following subscripts, other than those designating variables, a correlation coefficient with various corrections and under special conditions.

r,- a ratio.

r, or i,- an interest rate.

r_{12},- may be called a total correlation coefficient to distinguish it from partial correlation coefficients.

$r_{12.3}$, $r_{12.34}$, $r_{12.345}$, etc., are partial correlation coefficients.

$r_{1\Delta 23}$, $r_{1\Delta 234}$, etc., as herein used, designate multiple correlation coefficients. See Ch. XI, Sec. 3.

$r_{1.23}$, $r_{1.234}$, etc., have been used to designate multiple correlation coefficients. Such usage is avoided in this text for reasons given in Ch. XI, Sec. 3.

$r_{1(23)}$, $r_{1(234)}$, etc.,- an alternative notation to $r_{1\Delta 23}$, $r_{1\Delta 234}$, etc.

$R_{1(23)}$, $R_{1(234)}$, etc., are an alternative notation for $r_{1\Delta 23}$, $r_{1\Delta 234}$, etc.

ρ,- designates the correlation coefficient, based upon the squares of differences in rank.

s, S, σ, and Σ

s,- R. A. Fisher uses s to indicate an unbiased estimate of σ, which, in his usage, is a population value. Thus $s^2 = N/(N-1) \times V$ as used herein.

s,- as a preceding subscript to r, or R, indicates a correction for shrinkage.

s,- the indifferent value of a proportion, or of a mean, which is also the slope of boundary lines in sequential analysis.

S,- as a symbol of operation, indicates a summation. As used herein, Σ is a summation with reference to N, the cases in the sample, and S a summation with reference to the number of classes, or the like.

σ,- the standard deviation, which equals the square root of the variance. As used by R. A. Fisher, σ is a population, not a sample, value.

Σ,- as a symbol of operation it indicates a summation. See *S*. In this text Σ indicates a summation of N terms unless specially noted, thus $\sum^{N(N-1)}$ is a summation of $N(N-1)$ terms, and $\sum_{i=1}^{k}$ is a summation as i takes all values from 1 to k. This is also commonly written Σ_1^k. In this text the still more abridged notation Σ^k is sometimes used. Great care is necessary with double and triple summations. $\sum_{i=1}^{u} \sum_{j=1}^{v} A_{1i}A_{2j}a_{2i}a_{1j}$ is a summation of uv terms in which 1 and 2 do not vary and i takes all values from 1 to u simultaneously in A_{1i} and a_{2i}, and for each value of i all values of j from 1 to v (including the case in which $j = i$) simultaneously in A_{2j} and a_{1j}. If case $i = j$ is to be excluded, it must be indicated, usually by notation at the right of the equation, $j \neq i$. Another common situation is when j takes all values less than i, in which case the notation $j<i$ is recorded. In this case a double summation $\sum_{i=1}^{u} \sum_{j=1}^{u-1}$, $j<i$, would have $u(u-1)/2$ terms. Summations are sometimes indicated in mathematical treatises (seldom in statistical) by the use of a dummy index, subscript or superscript. When this index appears twice in an expression, it represents a summation, thus $a_{ij}\ b_{ik} = a_{1j}\ a_{1k} + a_{2j}a_{2k} + \ldots + a_{nj}a_{nk}$, in which 1, 2, ... n are all the possible values that i can take.

Σ,- the standard deviation. When two standard deviations, not readily definable by different subscripts, are involved, σ and Σ may be used. E.g., σ designates a standard deviation in a narrow range sample and Σ designates it in a wide range sample.

t, T, τ, θ, and Θ

t,- "Student's" t is a critical ratio,—a deviation in terms of its standard deviation. The t-distribution is symmetrical, but non-normal, because, and only because, the number of cases in the sample is small.

$t^2 = F_{1j}$,—see F_{ij}.

t,- a tabled entry. Ch. XIII, Sec. 12, Table XIII I.

θ,- a common designation for an angle.

θ_i,- as used herein, θ_i is the function depending on i, the d.o.f., involved in normalizing a variance ratio. See Ch. IX, Sec. 5. [9:23]

u, U, ν, and Υ

u,- commonly used to indicate a tabled entry.—see "t".

v and V

v,- in connection with percentiles, v_p is the value of the lower limit of the interval in which the (100p) percentile lies.

V,- the variance: V_1 is the variance of the first variable, V_2 that of the second, etc. For exception, see V_i and V_j.

V_i and V_j,- as here used in connection with variance ratios, V_i indicates a variance vased upon i d.o.f. and V_j one based upon j d.o.f.

V_{ij},- in connection with a normal distribution, is the variance of the portion of a unit normal distribution lying between x_i and x_j. See [8:29].

w and W

w,- designates a weight, e.g., w_1, w_2, w_3 may designate the weights attaching to the first, second, and third variables.

x, X, ξ and Ξ

x,- a score as a deviation from the mean.

x,- not infrequently a standard score, as in Table XV C.

x,- not infrequently a raw score, i.e., it $= X$ as herein used.

X,- a raw score.

$\overline{X}$,- the mean of the X's.

ξ,- a score as a deviation from an arbitrary origin.

y and Y

Same meanings for y, Y, and ζ, but pertaining to a second variable, as given for x, X, and ξ as pertaining to a first variable.

$Y_m(x)$,- a common designation of a Bessel function of the second kind.

z, Z, and ζ

z,- not infrequently a standard score.

z,- the ordinate in a unit normal distribution.

z,- R. A. Fisher's z. See Ch. XV, Sec. 1, 1938, Fisher and Yates.

ζ,- see "y".

ϕ and Φ

ϕ,- defined under f.

χ

χ,- the square root of χ^2. χ^2, though frequently derivable from categorical data and frequencies in classes, is, with varying degrees of closeness, the sum of a number of independent squared deviates of a unit normal distribution. $(\chi^2/i) = F_{i\infty}$, — see F_{ij}.

As herein used, i designates the number of degrees of freedom in χ_i^2.

ω and Ω

ω,- as a subscript is used in connection with corrections for attenuation to indicate a true score in the second variable. ∞ indicates the true score in the first variable.

SECTION 2. CERTAIN LESS COMMON NON-LITERAL SYMBOLS

$\cong$, $\doteq$, or $\fallingdotseq$, is approximately equal to.

$\sim$,- is asymptotically equal to. Frequently, in statistical equations, the nearness to equality increases as N increases.

$\Leftrightarrow$,- is equivalent to, in terms of the designated function relationship; thus if two achievement tests are equated on the basis of age norms, we might write $X = 78 \Leftrightarrow Y = 152$, meaning that 78 is the same age norm on test X as is 152 on test Y.

$\propto$,- varies as.

$>$,- is greater than.

$<$,- is less than.

$\geqq$,- is greater than, or in the limiting case, is equal to.

$\leqq$,- is less than, or in the limiting case, equal to.

$|x|$,- the absolute value of x, or the value regardless of sign.

$x!$, or $\underline{x|}$,- factorial x. If used when x is not an integer, it can then represent $\Gamma(x+1)$, but if so used, a specific statement in the accompanying text is necessary.

Γx,- see under Greek letter gamma.

$A \cdot B$,- the product of A and B. Also, in dealing with vectors, this is the "dot product," or scalar product of A and B.

$|a_{ij}|$,- the determinant whose element in the i-th row and j-th column is a_{ij}.

$\|a_{ij}\|$,- the matrix whose element in the ith row and jth column is a_{ij}.

$f(x)\Big]_a^b = f(b) - f(a)$

$|abc\ldots|$ is the determinant $\begin{vmatrix} a_1 & a_2 & a_3 & \cdots \\ b_1 & b_2 & b_3 & \cdots \\ c_1 & c_2 & c_3 & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{vmatrix}$

$\|abc\ldots\|$ is the matrix with the same elements as in the determinant $|abc\ldots|$.

$\overline{a+b} = (a+b)$

∞,- as a subscript in connection with corrections for attenuation, is used to indicate a score in the first variable having no chance error. ω is similarly used with reference to the second variable.

$\rightarrow$,- approaches, thus $n \rightarrow \infty$ is read, "as n approaches infinity."

$)i($,- reversed parentheses in a sequence written $1,2\ldots)i(\ldots,k$ asserts that all the values, 1, 2, etc., to k, except the value i, are included. E.g., should i have the specific value 2, the sequence would be $1,3,4,\ldots,k$. Where no ambiguity results, the commas may be omitted and this last series written $134\ldots k$.

[ ],- not infrequently square brackets are used to designate the mean of whatever is within the brackets. E.g., for a sample of N, $[X] = \Sigma X/N = M$. $[x_1x_2] = (\Sigma x_1x_2)/N = c_{12}$. Textual explanation of this use of [ ] is necessary, for otherwise it may be interpreted as a symbol of aggregation.

SECTION 3. A FEW MATHEMATICAL TERMS (not elsewhere defined) COMMONLY HAVING STATISTICAL SIGNIFICANCE

Affine,- an affine transformation is of the form

$$y_1 = a x_1 + b x_2 + c$$
$$y_2 = A x_1 + B x_2 + C$$

When $\begin{vmatrix} a & b \\ A & B \end{vmatrix} = (aB - Ab) \neq 0$, this covers the fundamental types of general utility in statistics: translations, rotations, stretchings and shrinkings, and reflections. The affine transformation carries parallel lines into parallel lines and finite points into finite points. The rotation is an isogonal affine transformation in that it does not change the size of angles.

Bernoullian numbers,- they are: $B_1 = \frac{1}{6}$; $B_2 = \frac{1}{30}$;

$B_3 = \frac{1}{42}$; $B_4 = \frac{1}{30}$; $B_5 = \frac{5}{66}$; $B_6 = \frac{691}{2730}$; $B_7 = \frac{7}{6}$;

$B_8 = \frac{3617}{510}$; . . . $B_n = \frac{(2n)!}{2^{2n-1}\pi^{2n}} \sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^{2n}$

Combination,- the number of possible combinations of n different things m at a time is designated $C(n,m)$, or ${}_nC_m$, or C^n_m, or $\binom{n}{m}$. $C(n,m) = N!/(n-m)!\,m!$ See *Permutation*.

Factor,- a factor as used in mental factor analysis, is a linear function of the scores upon a number of measures (usually mental tests) which, in relation to other factors, i.e., other linear functions of the same scores, has certain unique properties. Different con-

siderations lead to different factorial solutions. Involved are concepts of (*a*) independence, orthogonality and maximal variance (Hotelling, Kelley, et al.), (*b*) simple structure (Thurstone, et al.), (*c*) social importance (Kelley), and various hypotheses, general (Spearman, et al.), group (Thomson, et al.), bi-factor (Holzinger, et al.) etc., which qualify the conditions which are imposed.

Geometric series,- one in which the ratio of each term to the following is constant. Let the geometric series be $a + ar + ar^2 + \ldots + ar^n$. The sum of these ($n$+1) terms is $\frac{a(1 - r^{n+1})}{1 - r}$.

Integral function,- one which can be so written that the variable, or variables, have positive exponents and do not appear in the denominator of any term.

Integration,- is the process of summing parts (the infinitesimal elements of area of the calculus) to obtain a whole. An integral is the result of an integration. A definite integral is the result of integration between definite, or assigned, values.

Lexian ratio,- designated Q by Lexis $= \chi^2/\text{d.o.f.}$

Logarithm,- if given $y = a^x$, then x is the logarithm of y to the base a. Two bases are in common use,—the base e and the base 10, yielding natural logarithms designated $\ln$, or $\log_e$, and common logarithms designated log, or $\log_{10}$. With base e we have $y = e^x$ and the important property $\frac{dy}{dx} = e^x$ holds. With base 10 we have $y = 10^x$ and the important property $\log(10^p y) = p + \log y$ holds, thus $\log 172. = 2 + \log 1.72 = 2.235528\ldots$ in which 2 is called the charac-

teristic and .235528... the mantissa. p is a positive or negative integer.

The relationship between natural (or Napierian) and common (or Briggs) logarithms:

$$\log x = .43429\ 44819\ 03252\ \ldots\ \ln x$$

$$\ln x = 2.30258\ 50929\ 94046\ \ldots\ \log x$$

Orthogonal Transformation,- one which transforms from one set of rectangular coordinates to a second set. E.g., simple rotations and translations.

Permutation,- the permutations of n different things m at a time consist of all possible combinations m at a time and for those in each combination all possible arrangements. The number of such is designated $P(n,m)$, or ${}_nP_m$, or P^n_m. $P(n,m) = n!/(n-m)!$ See combination.

Quadrature,- the process of finding a rectangular area equal to that of a given surface. Usually one only of the bounds of this surface is curved.

Ratio test for convergence of an infinite series,- if the absolute value of the ratio of the $(n+1)$th term to the nth term as $n\to\infty$ is less than 1, the series is convergent; if equal to 1, there is no test; and if greater than 1, the series is divergent.

Rational function,- one which can be written so that the variable, or variables, do not have fractional exponents, or equivalent radical form.

Rational integral function,- one which can be so written as to be both rational and integral.

Vector,- a line having both magnitude and direction.

SECTION 4. KEY TO FORMULAS IN ORDER OF OCCURRENCE

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
Re percentiles: p, P_p, v_p, f_p, i_p, and F_p		146
P_p	[4:01]	146
i'_p and f'_p		149
σ_{P_p} and Q	[4:03]	149
A generalized mean	[6:01]	201
V_d	[6:03]	203
	[6:07]	204
V	[6:04]	203
	[6:05]	203
	[6:46]	217
σ	[6:06]	203
	[6:44]	217
$\widetilde{\widetilde{V}}$	[6:09]	204
$\tilde{V}$ and $\bar{V}$	[6:10]	205
$\widetilde{\widetilde{M}}$	[6:11]	206
$\tilde{X}$	[6:14]	208
V_M	[6:15]	209
	[6:16]	209
σ_M	[6:18]	209
$V(M_1 - M_2)$	[6:19]	210
Moment from fixed point	[6:20]	210

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
Mo_ξ	[7:24]	261
$V(Mo_\xi)$	[7:27]	262
V_{Mo}	[7:28]	262
V_f	[7:29]	265
Geometric mean	[7:30]	266
Harmonic mean	[7:31]	272
Differential equation of normal curve	[8:01]	288
z,- unit normal distribution	[8:02] [10:19]	288 348
p,- in a normal distribution	[8:03]	288
y,- normal distribution	[8:04] [8:34]	289 301
μ_1,- normal	[8:06]	290
A generalized mean	[7:01]	234
σ_{Mdn}	[7:02]	240
β_2,- Pearson	[7:03]	244
Equation of distribution for which $\sigma_M = \sigma_{Mdn}$	[7:05]	245
Ku,- percentile measure of kurtosis	[7:08]	246
V_{Ku}	[7:09]	246
Mesokurtic Ku	[7:11]	246
β_1,- Pearson	[7:12]	249
g_1,- Fisher	[7:13]	249
V_{g_1}	[7:14]	249

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
V_{k_3}	[7: 15]	250
A_s,- percentile measure of asymmetry	[7: 16]	250
V_{A_s}	[7: 17]	250
Sk,- percentile measure of skewness	[7: 18]	250
V_{Sk}	[7: 19]	251
Mean deviation,- normal	[8:07] [8: 20]	290 292
$\mu_2 = V$,- normal	[8:08]	290
Mean $\lvert x^3 \rvert$,- normal	[8:09]	291
μ_4,- normal	[8: 10]	291
Mean $\lvert x^5 \rvert$,- normal	[8: 11]	291
μ_n,- normal	[8: 12]	291
μ_6,- normal	[8: 13]	291
μ_8,- normal	[8: 14]	291
μ_{10},- normal	[8: 15]	291
β_1,- normal (Pearson)	[8: 16]	291
β_2,- normal (Pearson)	[8: 17]	291
γ_1,- normal (Fisher)	[8: 18]	291
γ_2,- normal (Fisher)	[8: 19]	291
Q,- normal quartile deviation	[8: 21]	292
Pv,- normal	[8: 22]	292
x (or z),- a standard score	[8: 23]	293

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
χ^2 and contingency table notation,-	[9:07] to [9:11]	318
f_s, $\tilde{f}_s$, χ^2_s, χ^2, d.o.f., $\tilde{f}_{st}$, f_{st}	[9:16] to [9:20]	325
Linear restriction	[9:12] to [9:13]	322
F_{ij},- variance ratio	[9:21]	326
f_{ij}	[9:22]	326
θ_i	[9:23]	326
d normalizing transformation	[9:24]	327
x normalizing transformation	[9:25]	327
P_{ij} and P_{ji}	[9:26]	329
Additive property of χ^2's	[9:27]	331
$\overline{X}_0$	[10:01] [10:01a]	333 333
a_{01}	[10:02]	333
b_{01}	[10:03]	333
$\overline{x}_0$ and $\overline{x}_1$ (See also [11:97])	[10:04] [10:04a]	333 333
$\overline{z}_0$ and $\overline{z}_1$	[10:05] [10:05a]	333 333

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
Preceding subscript notation indicating corrections in correlation coefficients: a_0, a_1, c_0, c_1, c', c'_0, c'_1, s	Table X A	334
$\overline{\overline{X}}_0$,- quadric regression	[10:06]	335
r	[10:07]	339
Comparable measures	[10:08]	340
$\sigma_{1.2}$	[10:09] [10:24] [10:49]	342 350 364
y,- bivariate normal distribution	[10:10] [10:21]	343 348
Δ subscript notation (See also Chapter XI, Sec. 3, pp. 435-436)		344
V_2,- analysis of variance (See also [11:99])	[10:11] [10:16] [10:17]	345 346 347
$V_{2.1}$	[10:12] [10:18]	346 347
$V_{2\Delta 1}$	[10:15]	346
$x_{2.1}$	[10:23]	349
$\sigma_{2.1}$	[10:09] [10:24] [10:49]	342 350 364

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
$\overline{X}$	[10:71]	372
$\overline{Y}$	[10:72]	372
$F_{1,N-2}$ for the difference between means of the upper and lower classes of the dichotomy	[10:73]	372
$b_{xy'}$,- y' assumed continuous	[10:74]	374
$r_{xy'}$,- biserial r	[10:75]	374
$\overline{y'}$	[10:76]	374
$V_{y'.x}$	[10:77]	375
V (biserial r)	[10:78] [10:79]	375 375
α, β, γ, δ, p, q, p', q' in 2X2-fold defined c_{xy},- 2X2-fold	[10:80]	380
ϕ,- product moment r in 2X2-fold	[10:81]	381
b_{xy}	[10:82]	381
$\overline{X}$	[10:83]	381
$F_{1(N-2)}$ to test $(p-\tilde{p})$	[10:84]	381
$F_{1(N-2)}$ to test $(b_{xy}-\tilde{b}_{xy})$	[10:85]	381
V_{ϕ}	[10:86]	382
r_t,- tetrachoric r	[10:87]	383

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
$V(r_t)$	[10:88]	386
r_t ,- when dichotomic lines are at the medians	[10:89]	386
	[10:89a]	387
$V(r_t)$	[10:90]	387
$r_{\tilde{x}\,x_m}$,- r between class means and continuous variate	[10:91]	393
${}_c{}'V_{x_m}$	[10:92]	393
${}_m r_{x_m y_m}$	[10:93]	393
	[10:94]	393
${}_c{}'C_{x_m y_m}$	[10:95]	393
X_α ,- a weighted sum	[10:96a]	395
	[10:96b]	395
V_α	[10:97]	395
$C_{\alpha\beta}$	[10:98]	395
$r_{\alpha\beta}$	[10:99]	395
	[10:100]	396
	[10:101]	396
$\bar{r}_{ig}$, $\bar{r}_{ij}$, $\bar{r}_{gh}$ average intercorrelations	[10:100]	396
$\bar{\rho}_{ij}$	[10:102]	397
	[10:105]	397
$\bar{\rho}_{01}$	[10:103]	397
V_d ,- variance of a difference	[10:106]	398

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
t_{1234}	[11:41]	421
$V(t_{1234})$	[11:42]	422
f_{12345}	[11:45]	424
Best estimate of z_{ω}	[11:44]	424
Weighting to allow for reliability	[11:45]	424
Effect of range upon reliability	[11:46a]	427
$v_{\omega .} / V_{\dot{\omega}}$	[11:47]	427
R_2	[11:48]	429
$v_{1.2} = V_{1.2}$	[11:49]	430
$b_{12} = B_{12}$	[11:50]	430
Effect of an imposed v_2/V_2 upon r_{12}	[11:51] [11:52]	430 430
Effect of normal selective processes upon r_{12}	[11:53] [11:55]	432 432
$\overline{X}_0$ in a 3-variable problem	[11:56] [11:56a]	433 433
$\overline{z}_0$	[11:57]	434
b_1 and b_2	[11:58]	434
a,- the constant term	[11:59]	434
$k_{0.12}$,- a 3-variable multiple alienation coefficient	[11:60] [11:62] [11:81]	434 435 441

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
β_1 and β_2,- standard score regression coefficients	[11:61]	435
	[11:62]	435
	[11:82]	441
	[11:83]	441
$V(z_0)$,- analysis of variance	[11:63]	436
$r_{0\Delta 12}$	[11:64]	436
	[11:65]	436
	[11:94]	445
$V_{0.12}$	[11:66]	436
V_0,- analysis of variance	[11:67]	437
d.o.f.	[11:68]	437
$F_{2,N-3}$ to test $(r_{0\Delta 12} - 0)$	[11:69]	437
V_0,- analysis of variance when $r_{12} = 0$	[11:70]	438
d.o.f.	[11:71]	438
$F_{1,N-3}$ to test $(b_1 - 0)$ when $r_{12} = 0$	[11:72]	438
$F_{1,N-3}$ to test $(b_1 - \tilde{b}_1)$ when $r_{12} = 0$	[11:73]	438
$r_{01.2}$	[11:74]	439
	[11:83]	441
	[11:95]	445
$\beta_1 = \beta_{01.2} =$ etc.	[11:75]	439
$b_1 = b_{01.2} =$ etc.	[11:76]	439
	[12:20]	461
$\overline{X}_{0.2}$	[11:77]	439

KEY PHRASE OR NUMBER	FORMULA NUMBER	PAGE
$F_{1,N-n-1}$,- to test $(b_i - \tilde{b}_i)$	[12:35]	474
${}_s r^2_{0\Delta 12\ldots n}$,- correction for shrinkage	[12:36]	474
$F_{n,N-n-1}$,- to test $(r_{0\Delta 12\ldots n}-0)$	[12:37]	475
$\mathcal{A}$,- augmented matrix	[12:38]	476
a^{ji},- inverse elements	[12:39]	476
$\mathcal{R}$,- predictor matrix	[12:40]	476
r^{ji},- inverse elements	[12:41]	477
$\mathcal{B}$,- regression coefficient matrix	[12:42] [12:44]	477 477
$\mathcal{r}$,- correlation coefficient matrix	[12:43]	477
$V(\beta_i)$ cf. [12:34]	[12:47] [12:48] [12:48a]	477 478 478
$V(\beta_i - \beta_j)$	[12:50]	478
$F_{1,N-n-1}$,- to test $(\beta_i - \beta_j)$	[12:51]	478
$r_{\beta_i\,\beta_j}$	[12:52]	478
$r_{b_i b_j}$	[12:53]	479
$X_{0\Delta 1,1^2,1^3}$,- cubic regression	[13:04]	483
$V_{0.1,1^2}$,- analysis of variance	[13:07]	488

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
d.o. f.	[13:06]	488
$F_{k-1,N-2-k}$,- to test $V(M_{0k})$	[13:08]	488
ϵ^2 after removal of linear trend	[13:10]	492
ϵ^2 after removal of quadric trend	[13:11]	492
Equivalent percentiles	[13:12]	500
Equivalent standard scores	[13:13]	500
Equivalent estimated true standard scores	[13:14]	501
Equivalent ratios	[13:15]	501
$V(\frac{X}{Y})$	[13:16]	504
V_Q'	[13:16a]	504
$\frac{w_1}{w_2}$,- optimal weight	[13:18]	506
M,- most reliable M	[13:19]	506
κ_1,- Pearson	[13:20]	508
κ_2,- Pearson	[13:21]	508
δ,- Craig	[13:22]	508
$\mu_{-\eta}$,- negative moment	[13:23]	510
Type VIII, IX, XI criterion	[13:24]	511
Type V criterion	[13:25]	512

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
$\frac{dy}{ydx}$,- Pearson curves	[13:27]	512
	[13:27a]	512
α_n,- Carver	[13:28]	513
α_n, recursion formula	[13:29]	513
Giving a, b_0, b_1, and b_2	[13:30] to [13:33]	514 514
y,- unit rectangular distribution	[13:35]	515
y,- unit parabolic distribution	[13:36]	515
y,- unit straight line distribution	[13:37]	515
y,- unit exponential distribution	[13:38]	515
y,- Type XII	[13:39]	516
y,- Type XIII	[13:41]	517
y,- Type II	[13:48]	518
y,- Type VII	[13:53]	518
Type VIII, IX, XI cubic in m	[13:58]	519
y,- Type VIII and IX	[13:59]	519
y,- Type XI	[13:60]	520
y,- Type V	[13:67]	521
y,- Type I	[13:73]	522
y,- Type VI	[13:74]	522
y,- Type IV	[13:75]	522

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
V_d, - V of a derived statistic	[13:76]	523
	[13:78]	524
	[13:80]	525
	[13:84]	526
V_σ	[13:77]	524
	[13:79]	525
	[13:81]	525
$r_{\omega\gamma}$, - r corrected for attenuation	[13:85]	527
	[11:25]	412
	[13:91]	529
$V(r_{\omega\gamma})$	[13:88]	528
$r_{1\gamma}$	[13:89]	529
$V(r_{1\gamma})$	[13:90]	529
i, - interval in a table	[13:92]	539
a's, t's, and Δ's in connection with tables	Table XIII I	540
p, - direct interpolation	[13:93]	539
p', - inverse interpolation	[13:94]	539
Δ and δ, - inverse interpolation	[13:95]	541
	[13:96]	541
t_p^{ii}, t_p^{iii}, t_p^{iv}	[13:97]	541
	[13:98]	541
	[13:99]	541
E^{ii}, E^{iii}, E^{iv}	[13:100]	542
	[13:101]	542
	[13:102]	542
a^{-ii}, a^{-iii}	[13:103]	543
	[13:104]	543

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
E^{-ii}, E^{-iii}	[13: 105]	543
	[13: 106]	543
V_p	[13: 107]	545
$\mu_3(p)$	[13: 108]	545
$\mu_4(p)$	[13: 109]	546
$b(f_a f_b)$	[13: 110]	546
$r(f_a f_b)$	[13: 111]	547
$c(f_a f_b)$	[13: 112]	547
$c(p_a p_b)$	[13: 113]	547
$c(f_{aa'} f_a)$	[13: 114]	547
$c(f_a f_{a'})$	[13: 115]	548
$c(M_1 f_{aa'})$	[13: 116]	548
$b(M_1 M_2)$	[13: 117]	548
$r(M_1 M_2)$	[13: 118]	548
$c(M_1 M_2)$	[13: 119]	548
$b(V_1 V_2)$	[13: 124]	550
$r(V_1 V_2)$	[13: 125]	550
$c(V_1 V_2)$	[13: 126]	550
	[13: 148]	555
$V(p_{ij})$	[13: 127]	551
$c(p_{gh} p_{ij})$	[13: 128]	551
$V(P_{ij})$	[13: 129]	552
$c(P_{gh} P_{ij})$	[13: 130]	552

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
$c\,(c_{12}c_{13})$	[13:152]	555
$c\,(c_{12}c_{34})$	[13:153]	555
h_1 and h_2,- intercepts in sequential analysis	[13:154]	559
	[13:155]	559
	[13:173]	568
	[13:174]	568
s	[13:156]	559
	[13:172]	568
L	[13:159]	560
	[13:183]	569
A	[13:160]	561
B	[13:161]	561
a	[13:162]	562
	[13:170]	568
b	[13:163]	562
	[13:171]	568
$\overline{n}_{p_1}$	[13:165]	564
$\overline{n}_s$	[13:169]	564
	[13:180]	569
L_M	[13:175]	568
t_1 and t_2	[13:176]	568
	[13:177]	568
L_s	[13:178]	569
$\overline{n}_M$	[13:179]	569
$\overline{n}_{M_1}$	[13:181]	569
$\mathfrak{a}$,- a matrix	[14:01]	573
$\mathfrak{a}'$	[14:02]	574

KEY PHRASE OR SYMBOL	FORMULA NUMBER	PAGE
$\mathcal{A}b$	[14:03]	574
$\mathcal{A}\mathcal{B}$	[14:04]	575
$\mathcal{I}$	[14:05]	575
$\mathcal{A}^{-1}$	[14:06]	576
$\mathcal{A}(\mathcal{B}+\mathcal{C})$	[14:08]	576
Scalar matrix	[14:09]	577
$\mathcal{A}\mathcal{B}\mathcal{C}$	[14:10]	577
A and Δ	[14:11] [14:16]	577 580
A_{ij} and Δ_{ij}	[14:12]	578
A expanded	[14:13]	579
A,- factors of	[14:14]	580
Simultaneous linear equations	[14:15]	580
Solutions	[14:17]	581
μ_{s+1} of point binomial	[14:18]	581
Moments	[14:19] to [14:25]	581 582
Poisson	[14:26]	582
Moments	[14:27] to [14:36]	582 583
Hypergeometric	[14:37]	584
Moments	[14:38]	585

SECTION 5. REFERENCES

Aitken, A. C., "Studies in practical mathematics I. The evaluation, with applications, of a certain triple product matrix," *Proc. roy. Soc. Edinb.*, 57:172-181 (1937).

Archibald, Raymond Claire, See Bateman, Harry.

Barlow's Tables of Squares, Cubes, Square-Roots, Cube-roots, and Reciprocals of All Integer Numbers up to 10,000 (London and New York: E. and F. N. Spon, Ltd., 1935).

Bartky, Walter, "Multiple sampling with constant probability," *Ann. math. Statist.*, 14:363-367 (1943).

Bartlett, M. S., "The square root transformation in the analysis of variance," *J. roy. Stat. Soc. Supp.*, 3:68-78 (1936).

Bateman, Harry, and Raymond Claire Archibald, "A guide to tables of Bessel functions," *Mathematical Tables and Other Aids to Computation*, vol. I (July, 1944).

Bernbaum, Z. W., "An inequality for Mill's ratio," *Ann. math. Statist.*, 13:245-246 (June, 1942).

Bowley, A. L., *Elements of Statistics*, Fifth Edition (New York: Charles Scribner's Sons, 1926).

Bregman, E. O., See Thorndike, E. L.

Brooks, Edith, See Karsten, Karl.

Brown, William, and Godfrey H. Thomson, *Essentials of Mental Measurement* (Cambridge: Cambridge University Press, 1921).

Burgess, James, "On the definite integral $\frac{2}{\sqrt{\pi}} e^{-t^2}$ with extended tables of values," *Trans. of the Roy. Soc. Edinb.*, 37:257-321 (1897-1898).

Camp, Burton H., "Probability integrals for a hypergeometric series," *Biom.*, 17:61-67 (1925).

———, "Probability integrals for the point binomial," *Biom.*, 16:163-171 (1924).

Chesire, Leone, Milton Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric*

Correlation Coefficient (Chicago: University of Chicago Bookstore, 1933).

Cochran, W. G., "The analysis of variance when experimental errors follow the Poisson or binomial laws," *Ann. math. Statist.*, 11:335-347 (1940).

——, "The χ^2 correction for continuity," *Iowa State College J. of Sci.*, 16:421-436 (1942).

Committee Approved by the Advisory Committee of Social and Economic Research in Agriculture, Report to the Social Science Research Council, "Collegiate mathematics needed in the social sciences," *Econometrica*, vol. I, no. 2.

Comrie, L. J., and H. O. Hartley, "Table of lagrangian coefficients for harmonic interpolation in certain tables of percentage points," *Biom.*, 32:183-186 (1941).

Cosens, C. R., "Notes on the computation of the Bessel function $I_n(X)$," *Mathematical Tables and Other Aids to Computation*, I:133-135 (January, 1944).

Cowden, Dudley J., "Correlation concepts and the Doolittle method," *J. Amer. statist. Ass.*, 38: 327-334 (September, 1943).

Craig, Cecil C., "A new exposition and chart for the Pearson system of frequency curves," *Ann. math. Statist.*, 7:16-28 (1936).

Crum, W. L., and A. C. Patton, *Economic Statistics* (Chicago and New York: A. W. Shaw Company, (1925).

David, F. N., *Tables of the Ordinates and Probability Integral of the Distribution of r in Small Samples* (Cambridge: Cambridge University Press (1938).

Davis, Frederick B., "A note on correcting reliability coefficients for range," *J. educ. Psychol.*, 35:500-502 (November, 1944).

Davis, Harold T., *Tables of the Higher Mathematical Functions* (Bloomington, Illinois: The Principia Press, vol. I, 1933; vol. II, 1935.

Day, Edmund E., "Standardization of the Construction of statistical tables," *Quart. Amer. statist. Ass.*, New Series No. 129, vol. 17 (1920).

DeMoivre, A., *Approximatio ad Summan Terminorum Binomii $(a+b)^n$ in Seriem Expansi.* (1733).

Doolittle, M. H., "Method employed in the solution of normal equations and the adjustment of a triangulation," *U. S. Coast and Geodetic Survey Report*, pages 115-120 (1878).

Dunlap, Jack W., and Albert K. Kurtz, *Handbook of Statistical Nomographs, Tables, and Formulas* (Yonkers-on-Hudson, New York: World Book Company, 1932).

Dwight, Herbert Bristol, *Tables of Integrals and Other Mathematical Data* (New York: Macmillan Company, 1934).

Dwyer, P. S., "The Doolittle technique," *Ann. math. Statist.*, 12:449-458 (December, 1941).

———, "The solution of simultaneous equations," *Psychometrika*, 6:101-129 (1941).

Elderton, W. Palin, *Notes on Statistical Processes*, "An alternative method of calculating the rough moments from the actual statistics," *Biom.*, 4:374-378 (1905-1906).

———, *Frequency Curves and Correlation* (London: Layton, 1906; Second Edition, 1927).

Ezekiel, Mordecai, See Tolley, H. R.

Fawcett, C. D., assisted by Alice Lee, etc., "A second study of the variation and correlation of the human skull, etc.," *Biom.*, I:408-467 (1902).

Feldman, Hyman M., "Mathematical expectation of product moments of samples drawn from a set of infinite populations," *Ann. math. Statist.*, 6: 30-52 (March, 1935).

Filon, L. N. G., and Karl Pearson, "On the probable errors of frequency constants and on the influence of random selection on variation and correlation," *Phil. Trans.*, 191:289-311 (1898).

Fisher, R. A., *Design of Experiments* (Edinburgh and London: Oliver and Boyd, Second Edition, 1937).

——, "The distribution of the partial correlation coefficient," *Metron,* 3:329-333 (1924).

——, "The goodness of fit of regression formulae and the distribution of regression coefficients," *J. roy. Statist. Soc.*, 85:597-612 (July, 1922).

——, "Influence of rainfall on the yield of wheat at Rothamstead," *Phil. Trans. roy. Soc. of London,* Series B, 213:89-142 (1923).

——, "Moments and product moments of sampling distributions," *Proc. Lond. Math. Soc.*, Series 2, 30:199-238 (December, 1928).

——, "On the interpretation of χ^2 from contingency tables and the calculation of *P*," *J. roy. Statist. Soc.*, 85:87-94 (1922).

——, "On the mathematical foundations of theoretical statistics," *Phil. Trans. roy. Soc. of Lond.*, Series A, 222:309-369 (1921).

——, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron,* 1:3-32 (1921).

——, *Statistical Methods for Research Workers* (Edinburgh and London: Oliver and Boyd, 1925 et. seq.).

Fisher, R. A., and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (London and Edinburgh: Oliver and Boyd, 1938; Second Edition, 1943).

Freeman, Harold, *Sequential Analysis of Statistical Data: Applications.* A report submitted by the Statistical Research Group, Columbia University, to the Applied Mathematics Panel, National Defense Research Committee (July, 1944).

Galton, Francis, "Family likeness in stature," *Proc. roy. Soc.*, XL:42 (1896).

——, "Regression towards mediocrity in hereditary stature," *J. Anthrop. Inst.* (1885).

Galton, Francis, "Typical laws of heredity," *J. roy. Inst.*, (February, 1877).

Garrett, Harry E., *Statistics in Psychology and Education* (New York and London: Longmans, Green and Company, 1926).

Glover, James W., *Tables of Applied Mathematics in Finance, Insurance, Statistics* (Ann Arbor, Michigan: George Wahr, 1923).

Hartley, H. O., See Pearson, Egon S.

Heron, David., See Pearson, Karl.

Hilferty, Margaret M., See Wilson, Edwin B.

Horn, Daniel, "Correction for the effect of tied ranks on the value of the rank difference correlation coefficient," *J. educ. Psychol.*, 33: 686-690 (December, 1942).

Hotelling, Harold, "Experimental determination of the maximum of a function," *Ann. math. Statist.*, 12:20-45 (March, 1941).

Isserlis, L., "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biom.*, 12:134-139 (1918).

———, "On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression," *Biom.*, 11: 50-86 (1916).

Jerome, Harry, *Statistical Method* (New York and London: Harper and Brothers, 1924).

Karsten, Karl and Edith Brooks, "'Retro' charts," *J. Amer. statist. Ass.*, 38:302-310 (1943).

Kelley, Truman L., *Crossroads in the Mind of Man* (Stanford University, California: Stanford University Press, 1928).

———, *Essential Traits of Mental Life* (Cambridge: Harvard University Press, 1935).

———, "The evidence of periodicity in short time series," *J. Amer. statist. Ass.*, 38:319-326 (1943).

———, "Individual testing with completion exercises," *Teach. Coll. Rec.*, 18:371-382 (1917).

Kelley, Truman L., *Interpretation of Educational Measurements* (Yonkers-on-Hudson, New York: World Book Company, 1927).

———, *The Kelley Statistical Tables*, Revision in press, 1947 (Cambridge: Harvard University Press).

———, "A new measure of dispersion," *Quart. Amer. statist. Ass.*, 743-749 (June, 1921).

———, "A new method for determining the significance of differences in intelligence and achievement scores," *J. educ. Psychol.*, 14:321-333 (September, 1923).

———, "The reliability coefficient," *Psychometrika*, 7:75-83 (June, 1942).

———, "The reliability of test scores," *J. educ. Research*, 3:370-379 (May, 1921).

———, "Ridge route norms," *Harvard Educational Review*, 10:309-314 (May, 1940).

———, "The selection of upper and lower groups for the validation of test items," *J. educ. Psychol.*, 30:17-24 (January, 1939).

———, *Statistical Method* (New York: Macmillan Company, 1924).

———, "An unbiased correlation ratio measure," *Proc. nat. Acad. Sci.*, 21:554-559 (September, 1935).

———, and Quinn McNemar, "Doolittle versus the Kelley-Salisbury iteration method for computing multiple regression coefficients," *J. Amer. statist. Ass.*, 24:164-169 (1929).

———, and F. S. Salisbury, "An iteration method for determining multiple correlation coefficients," *J. Amer. statist. Ass.*, 21:282-292 (1926).

Kendall, Maurice G., *The Advanced Theory of Statistics*, Volume I (London: J. B. Lippincott Company, 1943).

Kendall, Maurice G., See Yule, G. Udny.

Keynes, J. M., *A Treatise on Probability* (London: Macmillan and Company, 1921).

Kuder, G. F., and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika*, 2:151-160 (1937).

Kuder, G. F., See also Richardson, M. W.

Kurtz, Albert K., See Dunlap, Jack W.

Lee, Alice, See Fawcett, C. D.

Macdonnell, W. R., "On criminal anthropology and the identification of criminals," *Biom.*, I: 177-191 (1901).

McNemar, Quinn, See Kelley, Truman L.

Maher, Helen C., See Wilson, Edwin B.

Merrington, M., and Catherine M. Thompson, "Tables of percentage points of the inverted beta (F) distribution," *Biom.*, 33:73-88 (1943).

Mises, Richard von, *Probability, Statistics and Truth* (New York: The Macmillan Company, 1939).

Molina, Edward Charles Dixon, *Poisson's Exponential Binomial Limit* (New York: D. Van Nostrand Company, 1942).

National Bureau of Standards Tables, Prepared by the Federal Works Agency, Works Progress Administration, for the City of New York, Arnold N. Lowan, Technical Director (1939-1944).

Neyman, J., and Egon S. Pearson, "Contributions to the theory of testing statistical hypotheses" *Statistical Research Memoirs*, Vol. I (1936).

———, "Sufficient statistics and uniformly most powerful tests of statistical hypotheses," *Statistical Research Memoirs*, Vol. I (1936).

Patton, A. C., See Crum, W. L.

Pearson, Egon S., "The percentage limits for the distribution of range in samples from a normal population," *Biom.*, 24:404-417 (1932).

———, and H. O. Hartley, "The probability integral of the range in samples of *N* observations from a normal population," *Biom.*, 32:301-310 (1942).

Pearson, Karl, Editor, *Biometrika Publications* (London: University College, 1934-1938).

(London: University College, 1934-1938).

Pearson, Karl, Editor, *Cambridge University Press Tracts for Computers,* Department of Applied Statistics, University of London (London: Cambridge University Press, 1919-1937).

———, "Contributions to the mathematical theory of evolution," *Phil. Trans. roy. Soc.*, Series A, CLXXXV: 71ff (1894).

———, "Notes on the history of correlation," *Biom.*, 13:25-35 (1920).

———, "On an extension of the method of correlation by grades or ranks," *Biom.*, 10:416-418 (1914).

———, "On further methods of measuring correlation," *Mathematical Contributions to the Theory of Evolution, Biom. Lab. Publications*, University of London (London: Cambridge University Press (1907).

———, "On the correlation of characters not quantitatively measurable," *Phil. Trans. roy. Soc. Lond.*, Series A, CXCV: 1-47 (1900).

———, "On the influence of double selection on the variation and correlation of two characters, " *Biom.*, 6:111-112 (1908).

———, "On the influence of natural selection on the variability and correlation of organs," *Phil. Trans. roy. Soc. Lond.*, Series A, 200: 1-66 (1902).

———, "On the measurement of the influence of 'broad categories' on correlation," *Biom.*, 9: 116-139 (1939).

———, "On the moments of the hypergeometric series," *Biom.*, 16:157-162 (1924).

———, "On the probable error of a coefficient of correlation as found from a fourfold table," *Biom.*, 9:22-27 (1913).

———, "On the probable errors of frequency constants," *Biom.*, 9:1-10 (1913).

———, "On the probable errors of frequency constants," *Biom.*, 2:173-281 (1903).

Pearson, Karl, "On the systematic fitting of curves to observations and measurements," Part I, *Biom.*, 1:165-303 (1902).

———, "Skew variation in homogeneous material," *Phil. Trans. roy. Soc.*, Series A, CLXXXVI: 343ff (1893), and a supplement in *Phil. Trans. roy. Soc.*, Series A, CXCVII:443-459 (1901).

———, *Tables for Statisticians and Biometricians*, Part I (1914) and Part II (1931). (London: Cambridge University Press).

———, Editor, *Tables of the Incomplete Beta-Function* (Cambridge, England: The *Biometrika* Office, University College, 1934).

———, Editor, *Tables of the Incomplete Gamma-Function* (Cambridge, England: The *Biometrika* Office, 1934).

———, and David Heron, "On theories of association," *Biom.*, 9:159-315 (1913).

Pearson, Karl, See also Filon, L. N. G.

Peirce, B. O., *A Short Table of Integrals*, Second edition (1910); Third edition revised by W. F. Osgood (1929) (Cambridge: Ginn and Company).

Pepper, Joseph, "Studies in the theory of sampling," *Biom.*, 21:231-258 (1929).

Peters, C. C., and W. Van Voorhis, *Statistical Procedures and their Mathematical Bases* (New York and London: McGraw-Hill Book Company, 1940).

Powys, A. W., "Data for the problem of evolution in man," *Biom.*, I:30-49 (1901).

Quetelet, L. A. J., *Anthropometrie ou mesure des Differentes Facultes de l'Homme* (1871).

Read, Cecil B., "Centers of population of learned groups," *Science*, 90; no. 2375 (1939).

Richardson, M. W., and G. F. Kuder, "The calculation of test reliability coefficients based on the method of rational equivalence," *J. educ. Psychol.*, 30:681-687 (December, 1939).

Richardson, M. W., See also Kuder, G. F.

Rietz, Henry L., Editor, *Handbook of Mathematical Statistics* (Boston: Houghton Mifflin Company, 1924).

Robinson, G., See Whittaker, E. T.

Romanovsky, V., "Note on the moments of a binomial $(p+q)^n$ about its mean," *Biom.*, 15:410-412 (1923).

——, "On the moments of standard deviations and of correlation coefficients in samples from a normal population," *Metron*, vol. 5, no. 4:3-46 (1925).

Rulon, Phillip J., "Simplified procedure for determining the reliability of a test by split halves," *Harv. educ. Rev.*, 9:99-103 (January, 1939).

Saffir, Milton, See Chesire, Leone.

Salisbury, F. S., See Kelley, Truman L.

Salvosa, L. R., "Tables of Pearson's type III functions," *Ann. math. Statist.*, I, no. 2 and 3 (1930).

Scarborough, James B., *Numerical Mathematical Analysis* (Baltimore: The Johns Hopkins Press, 1930).

Secrist, Horace, *An Introduction to Statistical Methods* (New York: Macmillan Company, 1917).

Shen, Eugene, "The standard error of certain estimated coefficients of correlation," *J. educ. Psychol.*, 15:462-465 (1924).

——, "The reliability coefficient of personal ratings," *J. educ. Psychol.*, 16:232-236 (1925).

Sheppard, William Fleetwood, *The Probability Integral*, Completed and edited by the British Association for the Advancement of Science, Committee for the Calculation of Mathematical Tables (Cambridge, England: Cambridge University Press, 1939).

Smith, James G., and Acheson J. Duncan, *Sampling Statistics and Applications*, vol. II of *Fundamentals of the Theory of Statistics* (New York and London: McGraw-Hill Book Company, 1945).

Smithsonian Mathematical Tables, Hyperbolic Functions, Fifth reprint by George F. Becker and C. E. Van Orstand (Washington: Smithsonian Institute, 1942).

Snedecor, George W., *Statistical Methods* (Ames, Iowa: Collegiate Press, 1939).

Soper, H. E., "On the probable error of the biserial expression for the correlation coefficient," *Biom.*, 9:91-115 (1914).

Statistical Research Group, *Sequential Analysis of Statistical Data: Applications*, Prepared by the Statistical Research Group, Columbia University, for the Applied Mathematics Panel, NDRC. SRG Report 255 (New York: Columbia University Press, 1945).

Thomson, Godfrey H., See Brown, William.

Thompson, Catherine M., "Table of percentage points of the χ^2 distribution," *Biom.*, 32:187-191 (1941).

——, "Tables of percentage points of the incomplete beta-function," *Biom.*, 32:168-181 (1941).

Thompson, Catherine M., See also Merrington, M.

Thorndike, E. L., and E. O. Bregman, "On the form of distribution of intellect in the ninth grade," *J. educ. Res.*, 10:271-278 (1924).

Thurstone, L. L., *The Fundamentals of Statistics* (New York: Macmillan Company, 1925).

Thurstone, L. L., See also Chesire, Leone.

Tippett, L. H. C., "On the extreme individuals and range of samples taken from a normal population," *Biom.*, 17:364-387 (1925).

Tolley, H. R., and Mordecai Ezekiel, "The Doolittle method for solving multiple correlation equations versus the Kelley-Salisbury iteration method," *J. Amer. Statist. Ass.*, 22:497-500 (1927).

U. S. Bureau of Standards, *Tables of Probability Functions*, Prepared by the Federal Works Agency, Works Project Admin., for the State of New York (Washington: Vol. 1, 1941; Vol. 2, 1942).

Van Voorhis, W., See Peters, C. C.

Wald, Abraham, *Sequential Analysis of Statistical Data: Theory.* A report submitted by the Statistical Research Group, Columbia University, to the Applied Mathematics Panel, NDRC. (New York: Columbia University Press, 1943).

——, *A General Method of Deriving the Operating Characteristics of any Sequential Probability Ratio Test.* A memorandum submitted to the Statistical Research Group, Columbia University (April, 1944).

——, "On cumulative sums of random variables," *Ann. math. Statist.*, 15:283-296 (Sept., 1944).

——, "Sequential tests of statistical hypotheses," *Ann. math. Statist.*, 16:117-186 (June, 1945).

Wald, Abraham, and J. Wolfowitz, "Statistical tests based on permutations of the observations," *Ann. math. Statist.*, 15:358-372 (Dec., 1944).

——, and ——, "Sampling inspection plans for continuous production which insure a prescribed limit on the outgoing quality," *Ann. math. Statist.*, 16:30-49 (March, 1945).

Walker, Helen M., *Mathematics Essential for Elementary Statistics* (New York: Henry Holt and Company, 1934).

Wherry, R. J., "A new formula for predicting the shrinkage of the coefficient of multiple correlation," *Ann. math. Statist.*, 2:440-457 (1939).

Whittaker, E. T., and G. Robinson, *The Calculus of Observations* (London, etc.: Blackie and Son, Ltd., 1926).

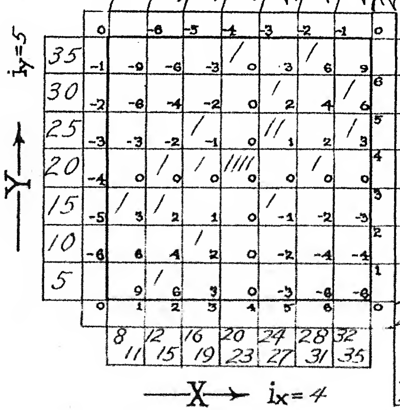
Wilson, Edwin B., Margaret M. Hilferty and Helen C. Maher, "Goodness of Fit," *J. Amer. statist. Ass.*, 26, New Series 176:443-448 (Dec., 1931).

Wishart, John, "The generalized product-moment distribution in samples from a normal multi-

variate population," *Biom.*, 20, Part A:32-52 (July, 1928).
Wolfowitz, J., See Wald, Abraham.
Yates, F., See Fisher, R. A.
Yerkes, Robert M., Editor, "Psychological Examining in the United States Army," *Nat. Acad. of Sci.*, Volume 15 (1921).
Young, Benjamin F., *Statistics as Applied to Business* (New York: Ronald Press, 1925).
Yule, G. Udny, "On the theory of correlation for any number of variables treated by a new system of notation," *Proc. roy. Soc. London*, A LXXIX:182-193 (1907).
———, *An Introduction to the theory of Statistics* (London and Philadelphia: Charles Griffin Company and J. B. Lippincott Company, 1912).
Yule, G. Udny, and M. G. Kendall, *An Introduction to the theory of Statistics* (London: Charles Griffin and Company, 1937).
Zizek, Franz, *Statistical Averages* (New York: Henry Holt and Company, 1913).

A chart of the type herewith, expanded to provide more than 12 classes in each dimension, is serviceable for manual or machine computation. (Illustrative computation shown in italics.)

Washington 6th Grade, — Lee Test
Variable X = Form 1 Scores
Variable Y = Form 2 Scores



f	1	3	3	5	4	2	2	20 N
ξ	-3	-2	-1	0	1	2	3	
$f\xi$	-3	-6	-3 (-12)		4	4	6 (14)	2 $\Sigma\xi$
$f\xi^2$	9	12	3		4	8	18	54 $\Sigma\xi^2$

Y	f	ζ	$f\zeta$	$f\zeta^2$
35	2	3	6	18
30	2	2	4	8
25	4	1	4	4
20	7	0	14	
15	3	-1	-3	3
10	1	-2	-2	4
5	1	-3	-3	9
	20		-8 / 6	46
	N		$\Sigma\zeta$	$\Sigma\zeta^2$

$i_y = 5$ — Y → —X→ $i_x = 4$

LL-UR diag.			
f	d	fd	fd²
	-6		
	-5		
	-4		
1	-3	-3	9
3	-2	-6	12
4	-1	-4	4
6	0	-13	
3	1	3	3
3	2	6	12
	3		
	4		
	5		
	6		
20		9 / -4	40
N		Σd	Σd^2

UL-LR diag.			
f	s	fs	fs²
	-6		
1	-5	-5	25
1	-4	-4	16
2	-3	-6	18
1	-2	-2	4
1	-1	-1	1
6	0	-18	
0	1	0	0
3	2	6	12
2	3	6	18
1	4	4	16
2	5	10	50
	6		
20		26 / 8	160
N		Σs	Σs^2

$\xi\zeta$ products by quadrants			
+−	−+	−−	++
-1	-1	3	6
		2	2
		2	6
		6	2
			3
-1	-1	13	19
-2		32	
30			
$\Sigma\xi\zeta$			

For summing diagonal f s cut out shaded portion

$V_\xi = \frac{\Sigma\xi^2}{N} - M_\xi^2 = .260$ $\quad \sigma_x = 1.6401 \quad \sigma_{\sigma_x} = .26$

$V_\zeta = \frac{\Sigma\zeta^2}{N} - M_\zeta^2 = 2.21$ $\quad \sigma_y = 1.4866 \quad \sigma_{\sigma_y} = .24$

Interpolate

$r = \dfrac{\frac{\Sigma\xi\zeta}{N} - M_\xi M_\zeta}{\sigma_\xi \sigma_\zeta} = .603$ $\quad M_x = 21.9 \quad \sigma_{M_x} = .37$

$\sigma_r = .15$ $\quad M_y = 21.5 \quad \sigma_{M_y} = .34$

Checks

$M_\xi = \frac{\Sigma\xi}{N} = .10 = \frac{\Sigma\zeta + \Sigma d}{2N} = .10$

$M_\zeta = \frac{\Sigma\zeta}{N} = .30 = \frac{\Sigma\zeta - \Sigma d}{2N} = .30$

$\Sigma\xi^2 = 54$

$\Sigma\zeta^2 = 46$

$\Sigma\xi^2 + \Sigma\zeta^2 = 100 = \frac{\Sigma\zeta^2 + \Sigma d^2}{2} = 100$

$\Sigma\xi\zeta = 30 = \frac{\Sigma s^2 - \Sigma d^2}{4} = 30$

INDEX